Virtual Symposium on the Virtual Mind

Stevan Harnad and other "Usenet" members

(Around 1992)

I thought you might find this interesting (and that you might even be tempted to join in). The discussion started with the following passage by Pat Hayes from a Virtual Symposium on the Virtual Mind that will appear in Minds & Machines in a few months. The rest is self-explanatory. I've included only the abstract plus the pertinent passages from the Symposium. (A few messages were unfortunately not saved, but I think they are easily reconstructed from context.)

-- Cheers, Stevan Harnad

[To Appear in: "Minds and Machines" 1992]

Virtual Symposium on the Virtual Mind

Patrick Hayes CSLI Stanford University

Stevan Harnad Psychology Department Princeton University

Donald Perlis Department of Computer Science University of Maryland

Ned Block Department of Philosophy and Linguistics Massachussetts Institute of Technology

ABSTRACT: When certain formal symbol systems (e.g., computer programs) are implemented as dynamic physical symbol systems (e.g., when they are run on a computer) their activity can be interpreted at higher levels (e.g., binary code can be interpreted as LISP, LISP code can be interpreted as English, and English can be interpreted as a meaningful conversation). These higher levels of interpretability are called "virtual" systems. If such a virtual system is interpretable as if it had a mind, is such a "virtual mind" real?

This is the question addressed in this "virtual" symposium, originally conducted electronically among four cognitive scientists: Donald Perlis, a computer scientist, argues that according to the computationalist thesis, virtual minds are real and hence Searle's Chinese Room Argument fails, because if Searle memorized and executed a program that could pass the Turing Test in Chinese he would have a second, virtual, Chinese-understanding mind of which he was unaware (as in multiple personality). Stevan Harnad, a psychologist, argues that Searle's Argument is valid, virtual minds are just hermeneutic overinterpretations, and symbols must be grounded in the real world of objects, not just the virtual world of interpretations. Computer scientist Patrick Hayes argues that Searle's Argument fails, but because Searle does not really implement the program: A real implementation must not be homuncular but mindless and mechanical, like a computer. Only then can it give rise to a mind at the virtual level. Philosopher Ned Block suggests that there is no reason a mindful implementation would not be a real one.

[text deleted]

HAYES: You have heard me make this distinction, Stevan (in the Symposium on Searle's Chinese Room Argument at the 16th Annual Meeting of the Society for Philosophy and Psychology in College Park, Maryland, June 1990). I now think that the answer is, No, Searle isn't a (possible) implementation of that algorithm. Let me start with the abacus, which is clearly not an implementation of anything. There is a mistake here (which is also made by Putnam (1975, p. 293) when he insists that a computer might be realized by human clerks; the same mistake is made by Searle (1990), more recently, when he claims that the wall behind his desk is a computer): Abacusses are passive. They can't actually run a program unless you somehow give them a motor and bead feelers, etc.; in other words, unless you make them into a computer! The idea of the implementation-independence of the computational level does not allow there to be NO implementation; it only suggests that how the program is implemented is not important for understanding what it does.

[text deleted]

Searle, J. R. (1990) Is the Brain a Digital Computer? Presidential Address. Proceedings of the American Philsophical Association.

> Date: Wed, 18 Mar 92 08:12:10 -0800
> From: searle@cogsci.Berkeley.EDU (John R. Searle)
> To: harnad@princeton.edu (Stevan Harnad) >
> Subject: Re: "My wall is a computer"
>
> Stevan, I don't actually say that. I say that on the standard Turing
> definition it is hard to see how to avoid the conclusion that
> everything is a computer under some description. I also say that I
> think this result can be avoided by introducing counterfactuals and
> causation into the definition of computation. I also claim that Brian
> Smith, Batali, etc. are working on a definition to avoid this result.
> But it is not my view that the wall behind me is a digital computer.
>
> I think the big problem is NOT universal realizability. That is only a
> SYMPTOM of the big problem. the big problem is : COMPUTATION IS AN
> OBSERVER RELATIVE FEATURE. Just as semantics is not intrinsic to syntax
> (as shown by the Chinese Room) so SYNTAX IS NOT INTRINSIC TO PHYSICS.
> The upshot is that the question : Is the wall (or the brain) a
> digital computer is meaningless, as it stands. If the question is "Can
> you assign a computational interpretation to the wall/brain?" the
> answer is trivially yes. you can assign an interpretation to anything.
>
> If the question is : "Is the wall/brain INTRINSICALLY a digital
> computer?" the answer is: NOTHING is intrisically a digital computer.
> Please explain this point to your colleagues. they seem to think the
> issue is universal realizability. Thus Chrisley's paper for example. >
> Anyhow the reference is to my APA presidential address " IS the Brain a

> Digital Computer?" proceeding of the Am Philos Assoc, for 90 or 91.
> I will send you the paper formatted for troff.
> Best john

John, many thanks for the reference and the details of your view about computers/computation. I think another way to phrase the question is:

(1) What is computation? and

(2) What is the implementation of a computation?

The answer I favor would be that computation is formal symbol manipulation (symbols are arbitrary objects that are manipulated on the basis of formal rules that operate only on their arbitrary shapes).

Syntax is unproblematic (just as it is in mathematics): It consists of rules that apply only to the arbitrary shapes of symbols (symbol tokens), not to their meanings. The problem is deciding what is NONTRIVIAL symbol manipulation (or nontrivially interpretable symbol manipulation): A symbol system with only two states, "0" and "1," respectively interpretable as "Life is like a bagel" and "Life is not like a bagel," is a trivial symbol system. Arithmetic and English are nontrivial symbol systems.

The trick will be to specify formally how to distinguish the trivial kind of symbol system from the nontrivial kind, and I suspect that this will turn out to depend on the property of systematicity: Trivial symbol systems have countless arbitrary "duals": You can swap the interpretations of their symbols and still come up with a coherent semantics (e.g., swap bagel and not-bagel above). Nontrivial symbol systems do not in general have coherently interpretable duals, or if they do, they are a few specific formally provable special cases (like the swappability of conjunction/negation and disjunction/negation in the propositional calculus). You cannot arbitrarily swap interpretations in general, in Arithmetic, English or LISP, and still expect the system to be able to bear the weight of a coherent systematic interpretation.

For example, in English try swapping the interpretations of true vs. false or even red vs. green, not to mention functors like if vs. not: the corpus of English utterances is no longer likely to be coherently interpretable under this arbitrary nonstandard interpretation; to make it so, EVERY symbol's interpretation would have to change in order to systematically adjust for the swap. It is this rigidity and uniqueness of the system with respect to the standard, "intended" interpretation that will, I think, distinguish nontrivial symbol systems from trivial ones. And although I'm not sure, I have an intuition that the difference will be an all-or-none one, rather than a matter of degree.

A computer, then, will be the physical implementation of a symbol system -- a dynamical system whose states and state-sequences are the interpretable objects (whereas in a static formal symbol system the objects are, say, just scratches on paper). A Turing Machine is an abstract idealization of the class of implementations of symbol systems; a digital computer is a concrete physical realization. I think a wall, for example, is only the implementation of a trivial computation, and hence if the nontrivial/trivial distinction can be formally worked out, a wall can be excluded from the class of computers (or included only as a trivial computer).

Best wishes, Stevan

Cc: Allen.Newell@cs.cmu.edu, GOLDFARB%unb.ca@UNBMVS1.csd.unb.ca (Lev Goldfarb), carroll@watson.ibm.com (John M Carroll), dennett@pearl.tufts.edu (Dan Dennett), fb0m+@andrew.cmu.edu (Frank Boyle), haugelan@unix.cis.pitt.edu, hayes@sumex-aim.stanford.edu (Pat Hayes), searle@cogsci.berkeley.edu

> Date: Fri, 20 Mar 92 8:47:13 EST
> From: Herb Simon
> To: Stevan Harnad
> Subject: Re: What is computation?
>
> A non-trivial symbol system is a symbol system that can be programmed to
> perform tasks whose performance by a human would be taken as evidence of
> intelligence.
>
> Herb Simon

Herb, this Turing-like criterion surely fits most cases of computation (though perhaps not all: we might not want to exclude mindless rote-iterations or tasks so unlike human ones that we might not even be able to say whether we would judge them as intelligent if performed by a human). But even if your criterion were extensionally equivalent to nontrivial computation, it still would not tell us what nontrivial computation was, because it does not tell us what "tasks whose performance by a human..." are! In other words, if this were the right criterion, it would not be explicated till we had a theory of what the human mind can do, and how.

In general, although the human element certainly enters our definition of computation (trivial and nontrivial) in that the symbol system must be systematically interpretable (by/to a human), I think that apart from that our definition must be independent of human considerations. I think it should be just as unnecessary to draw upon a theory of how the human mind works in order to explain what computation is as it is unnecessary to draw upon a theory of how the human mind works in order to explain what mathematics (or engineering, or physics) is.

Stevan Harnad

> Date: Fri, 20 Mar 92 08:56:00 EST
> From: "John M. Carroll"
> Subject: What is computation?
>
> Ref: Your note of Wed, 18 Mar 92 14:20:51 EST
>
> through a surprising quirk of intentionality on the part of the
> internet, i got copied on an exchange between you and john searle.
> thanks! it was related i think to my own ontological worries --
> you're interested in whether and how objects can be interpreted as
> computations or their implementations, i'm interested in whether
> and how designed artifacts can be interpreted as theories of their

> intended human users, or as implementations of those theories. since
> i've already eavesdropped, how did searle reply to your 'duals' idea?
> cheers

Hi John (Carroll)! You didn't eavesdrop; I branched it to you and others (by blind CC) intentionally, because I thought you might be interested. I've gotten several responses so far, but not yet from Searle. Dan Dennett wrote that he had published a similar "duals" idea, which he called the "cryptographer's criterion," and Frank Boyle wrote that Haugeland had made a similar rigid interpretability proposal in "AI and the Western Mind." I made the suggestion independently several years ago in a paper called "The Origin of Words: A Psychophysical Hypothesis" and first thought about it in reflecting on Quinean underdetermination of word meaning and inverted spectra several years earlier.

Although the artifact-design/user-theory problem and the problem of what is a computer/computation have some things in common, I suspect they part paths at the same Platonic point where the truths of formal mathematics part paths from the purposes of their creators. (Lev Goldfarb responded with a similar suggestion: that explaining nontrivial computation requires a theory of inductive learning.)

Stevan Harnad

Date: Sat, 21 Mar 92 03:14:43 EST To: roitblat@uhunix.uhcc.Hawaii.Edu (Herb Roitblat)

Herb (Roitblat), we disagree on a lot! I don't think a computer is the class of devices that can simulate other devices, or if it is, then that leaves me as uncertain what that class of devices is as before. I think a computer is a device that implements a nontrivial symbol system, and what makes a symbol system nontrivial is that it can bear the weight of one systematic interpretation (the standard one, and in a few special cases, some provable nonstandard ones). I think a grounded symbol system is one in which the interpretations of its symbols do not just square with what is in the mind of us outside interpreters, but also with what the system does in the real world. The nontrivial grounded symbol system that interests me is the robot that can pass the Total Turing Test (behave indistinguishably from ourselves).

We disagree even more on categories. I think the Roschian view you describe is all wrong, and that the "classical" view -- that categories have invariant features that allow us to categorize in the all-or-none way we clearly do -- is completely correct. Introspections about how we categorize are irrelevant (did we expect introspection to do our theoretical work for us, as cognitive theorists?), as are reaction times and typicality judgments. The performance capacity at issue is our capacity to learn to sort and label things as we do, not how fast we do it, not how typical we find the members we can correctly sort and label, not the cases we CANNOT sort and label, not the metaphysical status of the "correctness" (just its relation to the Skinnerian consequences of MIScategorization), and certainly not how we happen to think we do it. And the categories of interest are all-or-none categories like "bird," not graded ones like "big."

Cheers, Stevan

Date: Sun, 22 Mar 92 20:45:49 EST From: "Stevan Harnad" Subject: Re: What is computation? Cc: bradley@ivy (Bradley W Dickinson), briansmith.pa@xerox.com, dennett@pearl.tufts.edu (Dan Dennett), has@cs.cmu.edu, hayes@sumex-aim.stanford.edu (Pat Hayes), smk@wjh12.harvard.edu, sontag@gauss.rutgers.edu hatfield@linc.cis.upenn.edu, searle@cogsci.Berkeley.EDU

On: Nontrivial Computation, Nonarbitrary Interpretability, and Complexity

> Date: Sun, 22 Mar 92 17:07:49 -0500
> From: hatfield@linc.cis.upenn.edu (Gary Hatfield)
> To: harnad@Princeton.EDU, searle@cogsci.Berkeley.EDU
> Subject: Re: What is computation?
>
> Stevan: I don't see how you respond to John (Searle)'s point about observer-
> relativity. Indeed, your distinction between between trivial and
> nontrivial symbol systems appeals to an "intended interpretation,"
> which would seem simply to supply fuel for his fire. And your point
> about the wall being an implementation of a "merely trivial"
> computation is not clearly established: it depends on how you
> individuate the wall's computational states. John's claim (which is
> similar is some respects to one that Kosslyn and I made in our _Social
> Research_ paper, 1984, pp. 1025-6, 1032, and to Putnam's discussion in
> the appendix to his _Repn and Reality_) is that some state-assignment
> could be found for the wall in which it was performing any computation
> that you like, including any NONTRIVIAL computation (of course, the
> assignment might carve out physically arbitrary bits of the wall and
> count state transitions arbitrarily, from a physical point of view).
>
> Stephen and I, in our paper, contended that in order to avoid this sort
> of move the symbolist must argue that brains have non-arbitrary
> functional architectures, and that the functional architecture of our
> brain is so organized that it nonarbitrarily instantiates a serial
> digital computer. We then offered reasons for thinking that the brain
> doesn't have such an architecture.
>
> The crux of the matter is the status of function assignments. One might
> say that digital computers nonarbitrarily have a von Neumann
> functional architecture by offering a theory of the functions of
> artifacts according to which such functions are assigned relative to
> the intentions of designers or users. That might make sense of the
> intuition that commercial digital computers "intrinsically" are digital
> computers, though it wouldn't answer John's objection, because it still
> appeals to intentions in the function- assignment. But if one argued
> that there are natural functions, that biological systems
> nonarbitrarily instantiate one function rather than another, then the
> symbolists could claim (as I think Fodor and Pylyshyn did) that certain
> biological systems are naturally organized to compute with an
> architecture similar to that of digital comuters. John denies that

> there are natural functions. However, for his "observer-relativity" and
> "no intrinsic computations" arguments to have bite, he must do more
> than simply assert that there are no natural functions. Indeed, for the
> purpose of arguing against the computationalists, it would seem that he
> should offer them a choice between no-natural-functions and a trivial
> theory, and natural-functions but an empirically false theory.
>
> Best, Gary

Gary, thanks for your comments. Although I can't demonstrate it formally (but then a lot of this is informal and nondemonstrative), I suspect that there is a homology between a nonarbitrary sense in which a system is a computer and (the implementation of) a nontrivial computation, both resting on similar combinatorial, complexity-theoretic considerations. Coherent, systematic alternative interpretations are hard to come by, if at all, precisely because fitting an interpretation to a physical system is not arbitrary. There is, after all, a difference between a random string of symbols (typed by a chimp, say) that is (PERHAPS, and surely tortuously) interpretable as a Shakespearean play and a nonrandom string of symbols that is readily interpretable as a Shakespearean play. The complexity-theoretic difference would be that the algorithm you would need in order to interpret the random string as Shakespeare would be at least as long as the random string itself, whereas in the case of the real Shakespeare it would be orders of magnitude shorter. Moreover, one epsilon of perturbation in the random string, and you're back to square one insofar as its interpretability as Shakespeare is concerned. Not so with nonrandom strings and their interpretations. So interpretations the path to which is NP-complete hardly seem worth more attention than the possibility that this message could be interpreted as Grand Unified Field Theory.

I continue to think that we should be able to specify what (nontrivial) computation and computers are just as observer-independently as we can specify what flying, birds and aiplanes are. The only way the observer ever got into it in the first place was because a nontrivial symbol system must be able to bear the weight of a coherent systematic interpretation, which is something an observer might happen to want to project onto it.

Best wishes,

Stevan

> Date: Mon, 23 Mar 92 00:34:56 -0500
> From: hatfield@linc.cis.upenn.edu (Gary Hatfield)
>
> The claim that there are parsimonious and unpars. ways to "make sense"
> of input-output relations has long seemed to provide support for the
> belief that objective features of particular physical systems
> (computers, organisms) nonarbitrarily constrain content-ascriptions to
> those systems (e.g., Haugeland's Intro to _Mind Design_). The problem
> is that such arguments typically take as given some non-physical
> description of the inputs and outputs. Thus, Haugeland starts with
> "tokens," and in your reply to me you start with a "string of symbols"
> typed by a monkey. That doesn't speak to Searle's (or my) concerns, for

> one wants to know by what (non-observer-dependent) criteria you
> individuated the symbols. You need to argue that the physical state
> instantiating the symbols in the case of an actual Shakespeare play
> (including, say, the whole booklet; you aren't "given" symbol
> boundaries) has an internal coherence lacking in a given
> monkey-produced text *from a strictly physical point of view*. Here
> intuitions may diverge. But in any case, it seems to me that the
> defenders of non-trivial computation and non-arbitrary interpretation
> have the burden of starting their arguments from physical descriptions,
> without taking symbols for free.

Gary, Two-part reply: First, the bit-string generated by the black-white levels on the surface of the pages of a book look like a reasonable default encoding (then, together with a character-recognition algorithm and an English parser the string is parsimoniously reduced to a non-random one). But if that default option strikes you as too "observer-dependent," then pull the observer's MIND out of it entirely and simply allow the CAUSAL interaction -- between the book's surface optical structure (as demonstrably distinct from, say, the molecular structure of its ink) and organisms' visual transducers -- to serve as the objective basis for "picking out" the default encoding.

This uses only the fact that these symbols are parts of "dedicated" systems in the world -- not that any part of the system has a mind or interprets them -- in order to do the nonarbitrary parsing (the NP-completeness of "rival" reductions takes care of the rest). This is no different from the isolation of an experimental system in physics -- and it leaves computation as mind-independent as physics.

And if you rejoin that physics has the same "observer-dependence" problem (perhaps even citing quantum mechanical puzzles as evidence [which I would reject, by the way]), my reply is that computation is in good company then, and computers are no more or less of a natural kind than stones, birds or electrons.

Stevan Harnad

> Date: Sun, 22 Mar 92 21:58:42 HST
> From: Herbert Roitblat
>
> HR: You're right we do disagree on a lot. But, then, I knew that when I
> signed on to this discussion. What suprised me, however, is that I
> misunderstood what we disagreed about even more than I thought.
> I do not understand the following:
>
> >SH: implements a nontrivial symbol system, and what makes a symbol system
> >nontrivial is that it can bear the weight of one systematic
> >interpretation (the standard one, and in a few special cases, some
> >provable nonstandard ones).
>
> HR: I suspect you mean something special by "one systematic
> interpretation," but I do not know what you mean.

As an example, consider arithmetic, the scratches on paper, consisting of "0", "1", "+" etc., the axioms (strings of scratches) and rules of inference (applying to the scratches and strings of scratches). That’s a formal symbol system. The scratches on paper (symbol tokens) are manipulated only on the basis of their shapes, not what they "mean." (e.g., "0" is an arbitrary shape, and we have rules about what we can do with that shape, e.g., "0 + 1 = 1 + 0" etc.).

That’s the symbol system, and what we mean by numbers, equality, etc., is the systematic interpretation that can be PROJECTED onto those scratches on paper, and they will bear the weight of that interpretation. The very same scratches can also be given a few provably coherent "nonstandard" interpretations, but in general, rival interpretations simply won’t fit. For example, you cannot take the same set of scratches and interpret "=" as addition and "0" as equality and still come up with a coherent interpretation.

The same is true with the Sonnets of Shakespeare as a set of symbols interpretable as English, vs some other putative systematic interpretation of the very same scratches on paper.

> >SH: I think a grounded symbol system is one in which the interpretations
> >of its symbols do not just square with what is in the mind of us
> >outside interpreters, but also with what the system does in the real
> >world.
>
> HR: I agree with most of this, but I think that it does not matter whether
> the system’s symbols are interpretable or not, thus it does not matter
> whether they square with our expectations. I entirely endorse the
> idea that what is maximally important is what the system does.

It does matter for this discussion of what computation is, because computation is concerned only with systematically interpretable symbol systems, not random gibberish.

> >SH: The nontrivial grounded symbol system that interests me is the robot
> >that can pass the Total Turing Test (behave indistinguishably from
> >ourselves).
>
> HR: This is one source of our disagreement. I agree that the Turing test
> establishes a very high level of nontriviality, but I think that it is
> too high a level to be useful at this stage (a strategic issue) and
> is so high a level that it excludes much of what I find interesting.
> I would be happy with a system that MERELY (!?) passed the Turing test
> to the level of an ant or a rat or something like that. Why not just
> a gecko? I don’t think you mean that only humans are nontrivial
> computers. I cannot hope to live up to such standards in order to
> enter the discussion. I am still basically a comparative psychologist
> with interests in psycholinguistics.
>
> By the way, "trivial" is a conceptually dangerous term. When we fail
> to understand something it is nontrivial. Once we understand it, it
> becomes trivial.

There is more than one Turing Test (TT) at issue, and the differences between them are critical. The standard TT is purely symbolic (symbols in, symbols out) and calls for indistinguishability in all symbolic performance only. The Total Turing Test (TTT) I have proposed in its place (Harnad 1989, 1991) calls for indistinguishability in all symbolic AND robotic (sensorimotor interactions with the world of objects and events) performance. A lot rides on the TT vs. TTT distinction.

Nonhuman species TTT's would of course be welcome, and empirically prior to human TTT's, but unfortunately we lack both the ecological knowledge and the intuitive capacity (based on shared human homologies) to apply the TTT confidently to any species but our own. (This doesn't mean we can't try, of course, but that too is not what's at issue in this discussion, which is about what COMPUTATION is.)

I didn't say humans were computers, nontrivial or otherwise (they might be, but it seems to me they're also a lot of other things that are more relevant and informative). The question was about what COMPUTERS are. And I think "nontrivial" is a very useful term, a reasonable goal for discussion, and does not merely refer to what we have already understood.

> >SH: We disagree even more on categories. I think the Roschian view you
> >describe is all wrong, and that the "classical" view -- that categories
> >have invariant features that allow us to categorize in the all-or-none
> >way we clearly do -- is completely correct.
> >And the categories of interest
> >are all-or-none categories like "bird," not graded ones like "big."
>
> HR: This is a fundamental disagreement. It seems to me that your intent
> to focus on the most clearly classical cases derives from your belief
> that classical cases are the paradigm. Variability from the classical
> case is just "performance error" rather than competence. Am I correct
> on this last point?

Incorrect. I focus on categorical (all-or-none) categories because I think they, rather than graded categories, form the core of our conceptual repertoire as well as its foundations (grounding).

> HR: Bird is no less trivial than mammal, but we are faced with the
> question of whether monotremes are mammals. Living things are an all
> or none category. Are viruses living things? The question is not
> whether you believe viruses to be living things, you could be
> mistaken. Are they living things in the Platonic sense that classical
> theory requires? Bachelor is another classic category. Is a priest a
> bachelor? Is someone cohabiting (with POSSLQ) a bachelor? Is an 18
> year old unmarried male living alone a bachelor? Is a homosexual male
> a bachelor? What are the essential features of a bachelor and can you
> prove that someone either does or does not have them?

Herb, I've trodden this ground many times before. You just said before that you were a comparative psychologist. The ontology of the biosphere is hence presumably not your data domain, but rather the actual categorizing capacity and performance of human beings (and other species). It does not matter a whit to the explanation of the mechanisms of this performance capacity what the "truth" about montotremes, viruses or priests is. Either we CAN categorize them correctly (with respect to some Skinnerian consequence of MIScategorization, not with respect to some Platonic reality that is none of our business as psychologists) or we cannot. If we can, our success is all-or-none: We have not said that cows are 99% mammals whereas monotremes are 80% mammals. We have said that cows are mammals. And montotremes are whatever the biological specialists (hewing to their OWN, more sophisticated consequences of MIScategorization) tell us they are. And if we can't say whether a priest is or is not a bachelor, that too does not make "bachelor" a graded category. It just means we can't successfully categorize priests as bachelors or otherwise!

We're modelling the cognitive mechanisms underlying our actual categorization capacity; we're not trying to give an account of the true ontology of categories. Nor is it relevant that we cannot introspect and report the features (perfectly classical) that generate our success in categorization: Who ever promised that the subject's introspection would do the cognitive theorist's work for him? (These are all lingering symptoms of the confused Roschian legacy I have been inveighing against for years in my writings.)

> The classic conceptualization of concepts is tied closely to the
> notion of truth. Truth can be transmitted syntactically, but not
> inductively. If features X are the definition of bachelor, and if
> person Y has those features then person Y is a bachelor. One problem
> is to prove the truth of the premises. Do you agree that the symbol
> grounding problem has something to do with establishing the truth of
> the premises?

Nope. The symbol grounding problem is the problem that formal symbol systems do not contain their own meanings. They must be projected onto them by outside interpreters. My candidate solution is robotic grounding; there may be others. Leave formal truth to the philosophers and worry more about how organisms (and robots) actually manage to be able to do what they can do.

> The truth of the premises cannot be proved because we have no
> infallible inductive logic. We cannot prove them true because such
> proof depends on proving the truth of the implicit ceteris paribus
> clause, and just established that proof of a premise is not possible.
> We cannot be sure that our concepts are correct. We have no proof
> that any exemplar is a member of a category. I think that these
> arguments are familiar to you. The conclusion is that even classic
> categories have only variable-valued members, even they cannot truly
> be all-or-none.

The arguments are, unfortunately, familiar mumbo-jumbo to me. Forget about truth and ontology and return to the way organisms actually behave in the world (including what absolute discriminations they can and do make, and under what conditions): Successful (TTT-scale) models for THAT is what we're looking for. Induction and "ceteris paribus" has nothing to do with it!

> I think, therefore, that we are not justified in limiting discussion
> to only those categories that seem most clear, but that we would be
> served by developing a theory of conceptual representation that did
> not depend on artificial demarcations. I argue for a naturalistic
> theory of categories that depends on how people use conceptual labels
> (etc.). I argue that such use depends on a certain kind of

> computation, that given enough time, people could endorse a wide range
> of categorizations. The range of categorizations that they can
> endorse is the range of dimensions for which they represent the
> concept as having a value. My hunch is that the number of these
> dimensions that can be used at any one time for performing a given
> task is small relative to the number of dimensions that they know
> about for the concept. You seems more interested in characterizing
> the range of dimensions along which people can use their concept, I am
> more interested in the way in which they select those dimensions for
> use at the moment.

I'm interested in what mechanisms will actually generate the categorization capacity and performance of people (and animals). My own models happen to use neural nets to learn the invariants in the sensory projection of objects that will allow them to be categorized "correctly" (i.e., with respect to feedback from the consequences of MIScategorization). The "names" of these elementary sensorimotor categories are then grounded elementary symbols that can enter into higher-order combinations (symbolic representation), but inheriting the analog constraints of their grounding.

> Finally, I have been thinking about symbol grounding in other
> contexts. Exactly what symbols do you think are grounded in human
> representation? It cannot be letters because no semantics is
> attributed to them. It cannot be words, because we understand
> paraphrases to mean the same thing, the same word has multiple
> meanings, etc. It cannot be sentences, because we are productive in
> our use of sentences and could utter an indefinite number of them.
> The symbols would have to be internal, variably mappable onto surface
> symbols, and as such, not communicable with great confidence to other
> individuals. You would argue (yes?) that they are finite and discrete,
> but highly combinable. You would not argue, I think, that they get
> their meaning through their reference to some specifiable external
> object or event (i.e., you would not get into the Golden Mountain
> conundrum). Is symbol grounding nothing more than whatever relationship
> allows one to avoid unpleasant consequences of misunderstanding and
> misclassification (your allusion to Skinner)?

I don't know what the ground-level elementary symbols will turn out to be, I'm just betting they exist -- otherwise it's all hanging by a skyhook. Nor do I know the Golden Mountain conundrum, but I do know the putative "vanishing intersections" problem, according to which my approach to grounding is hopeless because not even sensory categories (not to mention abstract categories) HAVE any invariants at all: My reply is that this is not an apriori matter but an empirical one, and no one has yet tried to see whether bottom-up sensory grounding of a TTT-scale robot is possible. They've just consulted their own (and their subjects') introspections on the matter. I would say that our own success in categorization is some inductive ground for believing that our inputs are not too underdetermined to provide an invariant basis for that success, given a sufficiently powerful category learning mechanism.

> By the way, I am sorry for constantly putting words into
> your mouth, but it seems to me to be an efficient way to finding out
> what you mean.

In the event, it probably wasn't, but I managed to say what I meant anyway. I have an iron-clad policy of not sending people off to look up chapter and verse of what I've written on a topic under discussion; I willingly recreate it on-line from first principles, as long as my interlocutor does me the same courtesy -- and you haven't sent me off to chapter and verse either. I find this policy easy enough to be faithful to, because I don't have any ideas that cannot be explained in a few paragraphs (nothing longer than a 3-minute idea). Nor have I encountered many others who have longer ideas (though I have encountered many others who have been longer-winded or fuzzier about describing them).

> >SH: Introspections about how we categorize are irrelevant (did we expect
> >introspection to do our theoretical work for us, as cognitive
> >theorists?), as are reaction times and typicality judgments.
>
> HR Introspections play no role in my conceptualizaiton. You must have me
> confused with someone else. I am not even sure that I am conscious,
> let alone capable of introspection.

Regarding the non-role introspections play in your conceptualization, see what you asked me about the essential features of bachelors above. Why should I be able to introspect essential features, and what does it prove if I can't? All that matters is that I can actually sort and label bachelors as I do: Then finding the features I use become's the THEORIST's problem, not the subject's.

I would suggest, by the way, that you abandon your uncertainty about whether anybody's home inside you, experiencing experiences (as I confidently assume there is in you, and am certain there is in me). Cartesian reasons alone should be sufficient to persuade you that the very possibility of experiencing uncertainty about whether there is somebody home in your own case is self-contradictory, because "uncertainty" or "doubt" is itself a experiential state.

> If this discussion is heading off in a direction irrelevant to your
> interests, we can wait for another more opportune time. I think that
> our discssion has taken roughly this course: What is computation?
> Computation is either any regular state change (my position) or it is
> the set of nontrivial operations involving a grounded symbol set
> (fair?). What is a grounded symbol set?
>
> Aloha. Herb

The discussion is about what computers/computation are, and whether there is any principled way to distinguish them from what computers/computation aren't. In one view (not mine), what is a computer is just a matter of interpretation, hence everything is and isn't a computer depending on how you interpret it. In my view, one CAN distinguish computers -- at least those that do nontrivial computation -- on a complexity-theoretic basis, because systematic interpretations of arbitrary objects are as hard to come by as chimpanzees typing Shakespeare.

Now once we have settled on what computers/computation are (namely, nontrivial symbol manipulation systems), we still face the symbol grounding problem: These nontrivially interpretable systems still do not "contain" their own interpretations. The interpretations must be projected onto them by us. A grounded symbol system is one whose robotic performance in the real world of objects and events to which its symbols can be interpreted as referring squares systematically with the interpretation. The symbol interpretations are then grounded in its robotic performance capacity, not just in our projections.

References (nonobligatory) follow.

Cheers, Stevan

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Harnad, S. (1990a) The Symbol Grounding Problem. Physica D 42: 335-346.

Harnad, S. (1990b) Against Computational Hermeneutics. (Invited commentary of Eric Dietrich's Computationalism) Social Epistemology 4: 167-172.

Harnad, S. (1990c) Lost in the hermeneutic hall of mirrors. Invited Commentary on: Michael Dyer: Minds, Machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence 2: 321 - 327.

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on the Virtual Mind. Minds and Machines (in press)

Date: Sun, 29 Mar 92 18:39:31 EST From: "Stevan Harnad"

> Date: Mon, 23 Mar 1992 15:35:43 -0500 (EST)
> From: Franklin Boyle
> To: "Stevan Harnad"
> Subject: Re: What is computation?
>
> The article by Haugeland contains the following text which is relevant to the
> discussion, though it wasn't intended to address the SS issue at hand:
>
> Suppose that, instead of giving the analogy [interpretation],
> I had just spelled out the rules, and then invited you to
> discover the interpretation. That would be a cryptarithmetic
> puzzle, or, more generally, a code cracking assignment. The
> principle of all such efforts, from deciphering ancient inscript-

> tions to military cryptography, is finding a consistent reading
> such that the results reliably *make sense* [footnote to Quine].
> This requirement is by no means trivial or easy to meet; there
> are, for instance, no semantic interpretations attached to the
> chess or checkers systems. Hence, though an interpretation is
> never properly part of a formal system, the structure of a system
> strongly constrains possible interpretations; in other words, the
> relation between a formal system and a viable interpretation is
> not at all arbitrary.
> -- (p27, "Artificial Intelligence and the Western Mind.
> in _The Computer and the Brain: Perspectives on
> Human and Artificial Intelligence_, J.R. Brink &
> C.R. Haden (eds).)
>
> He is speculating about original meaning (rather than the difference
> between trivial and non-trivial SS's) and the fact that computers
> represent a technological solution to the "Paradox of Mechanical
> Reason" ["either meanings matter to the manipulations, in which case
> reasoning is not really mechanical (it presupposes an homunculus); or
> else meanings do not matter, in which case the machinations are not
> really rational (they are just some meaningless "machine-like"
> interactions" (p23)] because they "take care of the formalism" so that
> "any meaning that can be taken care of by taking care of the rules will
> be *automatically* taken care of by the computer -- without any paradox."
> (p27).
>
> -Frank Boyle

Frank, thanks for the passage. As I noted earlier, not only I, but also Dan Dennett came up with something like this independently. But I would stress that the uniqueness (or near-uniqueness, modulo duals) of the standard interpretation of a given symbol system, remarkable though it is (and this remarkable property is at the heart of all of formal mathematics), it is not enough to make that interpretation intrinsic to the system: If the right interpretation is projected onto the system, it will square systematically with the interpretation, but the projection will still be from outside the system. That's good enough for doing formal maths, but not enough for modelling a mind. For the latter you need AT LEAST a grounded symbol system.

Stevan Harnad

Date: Tue, 31 Mar 92 19:38:10 EST From: "Stevan Harnad" To: chrisley@oxford.ac.uk Subject: Re: What is computation?

> From: Ronald L Chrisley
> Date: Wed, 25 Mar 92 16:13:22 GMT
> To: harnad@Princeton.EDU
> Cc: chrisley@oxford.ac.uk, dave@cogsci.indiana.edu
>
> Stevan:
>

> RC: Here are some of my thoughts on the first part of your recent message.
> Could you provide the context for this exchange between you and
> Searle? Did this dialogue take place on the symbol-grounding list?

SH: Ron, no, the "What is Computation" discussion was actually initiated by a reply by John Searle to a passage from a 4-way "skywriting" exchange that will be published in the journal Minds and Machines under the title "Virtual Symposium on the Virtual Mind." The authors are Pat Hayes, Don Perlis, Ned Block and me. The passage in question was by Pat Hayes, in which he cited John Searle as claiming his wall was a computer.

I will send you the full exchange separately. Meanwhile, you wrote:

> > Date: Wed, 18 Mar 92 08:12:10 -0800
> > From: searle@cogsci.Berkeley.EDU (John R. Searle)
> > To: harnad@princeton.edu (Stevan Harnad)
> >
> > Subject: Re: "My wall is a computer"
> >
> > JS: Stevan, I don't actually say that. I say that on the standard Turing
> > definition it is hard to see how to avoid the conclusion that
> > everything is a computer under some description. I also say that I
>
> RC: No, actually Searle argues that the standard notion of computation
> implies that everything is *every* computer. Thus, he claims that his
> wall could be seen as implementing Wordstar. But of course, there are
> good reasons for ruling out such bizarre interpretations: for one,
> they're not causal.
>
> > JS: think this result can be avoided by introducing counterfactuals and
> > causation into the definition of computation. I also claim that Brian
> > Smith, Batali, etc. are working on a definition to avoid this result.
> > But it is not my view that the wall behind me is a digital computer.
>
> RC: Nor is it anyone else's view. That's because the standard view is
> that the physics *does* constrain computational interpretations. If
> it isn't the explicit standard view, it is implicit in the notion of a
> Turing *machine*. And if Searle wants to contest that it isn't even
> implicit, then his arguments only establish the superiority of a
> theory of computer science that is physically grounded, *not* the
> incoherence of the notion that a particular form of computation is the
> essence of mind.

SH: If I may interpolate some commentary: I agree about the physical grounding as picking out this machine running WORDSTAR as a privileged interpretation. I would add only two remarks.

(1) I think (though I can't prove it) that there is probably a complexity-based way of picking out the privileged interpretation of a system as a computer running a program (rather than other, more arbitrary interpretations) based on parsimony alone.

(2) This discussion of what a computer is does not necessarily have any bearing on the question of what the mind is, or whether the brain is a computer. One could argue yes or no that computers/computation pick out a nonarbitrary kind. And one can independently argue yes or no that this has any substantive bearing on what kind of system can have a mind. (E.g., I happen to agree with Searle that a system will not have a mind merely because it implements the right computer program -- because, according to me, it must also be robotically grounded in the world -- but I disagree that there is no nonarbitrary sense in which some systems are computers and others are not. I.e., I agree with him about [intrinsic] semantics but not about syntax.)

> > JS: I think the big problem is NOT universal realizability. That is
> > only a SYMPTOM of the big problem. the big problem is: COMPUTATION IS AN
> > OBSERVER RELATIVE FEATURE. Just as semantics is not intrinsic to syntax
> > (as shown by the Chinese Room) so SYNTAX IS NOT INTRINSIC TO PHYSICS.
> > The upshot is that the question : Is the wall (or the brain) a
> > digital computer is meaningless, as it stands. If the question is "Can
> > you assign a computational interpretation to the wall/brain?" the
> > answer is trivially yes. you can assign an interpretation to anything.
>
> RC: This kind of equivocation is the reason why I delineated 3 ways in
> which one might understand the claim "the brain is a computer".
>
> One is that it admits of any computational description at all. If
> this were the extent of the claim for cognitive science, then it is
> indeed unenlightening, since even a stone could be seen as being a
> Turing Machine with only one state.
>
> A second way of interpreting the claim is that there is a class of
> Turing Machine descriptions that are sufficiently complex that we
> would consider them as descriptions of computers, more conventionally
> understood, and that the brain, as opposed to a stone or Searle's wall
> (they just don't have the right properties of plasticity, input/output
> connections, causally related internal states, etc), admits of one of
> these descriptions.
>
> A third way of understanding the cognitivist claim is: the brain
> admits of a computational description, and anything that has a mind
> must also admit of a similar computational description. This is not
> vacuous, since most things, including not only stones and Searle's
> wall, but also bona fide computers, will not admit of such a
> description.

SH: It seems to me that everything admits of a trivial computational description. Only things with a certain kind of (not yet adequately specified) complexity admit of a nontrivial computational description (and those are computers). Now things that have minds will probably also admit of nontrivial computational descriptions, hence they too will be computers, but only in a trivial sense insofar as their MENTAL capacities are concerned, because they will not be ONLY computers, and their noncomputational robotic properties (e.g., transducers/actuators and other analog structures and processes) will turn out to be the critical ones for their mental powers; and those noncomputational properties will at the same time ground the semantics of the system's symbolic states.

> > JS: If the question is : "Is the wall/brain INTRINSICALLY a digital
> > computer?" the answer is: NOTHING is intrisically a digital computer.
> > Please explain this point to your colleagues. they seem to think the
> > issue is universal realizability. Thus Chrisley's paper for example.

SH: I unfortunately can't explain this for Searle, because I happen to disagree with him on this point, although I do recognize that no one has yet come up with a satisfactory, principled way of distinguishing computers from noncomputers...

> RC: I think that we can make sense of the notion of something
> intrinsically being a digital computer. Searle's argument that we
> cannot seems to be based on the claim that anything can be seen as a
> computer. In that sense, the issue for Searle *is* universal
> realizability. That is, Searle seems to be claiming that since the
> property *digital computer* can be realized by any physical system,
> then nothing is intrinsically a digital computer, and so viewing the
> brain as one will have little value.
>
> I disagree, of course, and on several counts. For one thing, on the
> second way of understanding the claim that something is a computer,
> the property *digital computer* is not universally realizable. But
> even if everything were *some* kind of digital computer (on the first
> or second ways of understanding), that would not invalidate the
> computational approach to understanding the mind, since that approach
> seeks to understand what *kind* of computation is characteristic of
> the mental (the third way). In fact, it would be of some use to
> cognitive science if Searle could show that everything is some kind of
> computer, because there are some critics of cognitive science who
> argue that the brain cannot be viewed as a computer at all (Penrose?).
>
> Searle's only options are to endorse the coherence of the cognitivist
> claim (I am not claiming that it has been shown to be true or false,
> just that it is coherent and non-trivial), find another argument for
> its incoherence, or deny my claims that causality is relevant to
> computational interpretation, thus suggesting that cognitivism is
> vacuous since every physical system can be interpreted as being every
> kind of computer. And even if he argues that causality is irrelevant
> to a *particular* style of computational interpretation, he has to
> show that it is irrelevant to any notion of computation before he can
> rule out any computational approach to mind as being incoherent. Put
> the other way around, he would have to show that a notion of
> computation that takes causality seriously would ipso facto not be a
> notion of computation. This seems impossible. So it looks like
> Searle must try to reject cognitivism some other way, or accept it.
>
> I tried to make all this clear in my paper. Due to publisher's

> delays, there are still chances for revisions, if anyone would like to
> suggest ways that I could make these points more clear.
>
> One last thing: given the reluctance that some AI/CompSci/CogSci
> people have to taking causality, connections to the world, etc.
> seriously, I welcome and encourage Searle’s points in some sense. I
> just wish he would see his arguments as establishing one type of
> cognitivism (embodied) to be prefereable to another (formal).
>
> Much of what people do in AI/CompSci/CogSci is the former, it’s just
> their theories of what they are doing that are the latter. I think
> the point of Searle’s paper is not "Cognitivism is incoherent" but
> rather "If you want to be a cognitivist, your theories better take
> seriously these notions of causality, connections to the world, etc.
> that are implicit in your practice anyway".
>
> Perhaps Searle’s points, cast in a different light, would not give
> people reason to abandon cognitivism, but would instead show them the
> way toward its successful development. As I said in my paper, "Searle
> has done us a service".
>
> Ronald L. Chrisley New College Oxford OX1 3BN

SH: I don’t think you’ll be able to get computer scientists or physicists excited about the factor of "causality" in the abstract, but IMPLEMENTATION is certainly something they think about and have views on, because a program is just as an abstraction until and unless it’s implemented (i.e., realized in a dynamical physical ["causal"] system -- a computer). But there’s still not much room for a convergence of views there, because good "symbolic functionalists" hold that all the particulars of implementation are irrelevant -- i.e., that the same program can be implemented in countless radically different ways with nothing in common except that they are all implementations of the same computer program. Hence the right level to talk about is again the purely symbolic (copmputational) one. I happen to disagree with these symbolic functionalists insofar as the mind is concerned, but not because I think there is something magic about the "causality" of implementation, but because I think a symbol system is just as ungrounded when it’s implemented as when it’s just scratches on static paper. The mere implementation of a program on a computer is the wrong kind of "causality" if a mind is what you’re interested in implementing (or even if it’s an airplane or a furnace). What’s needed is the robotic (TTT) power to ground the interpretations of its internal symbols in the robot’s interactions with the real world of objects, events and states of affairs that its symbols are interpretable as being "about" (TTT-indistinguishably from our own interactions with the world). (I list some of the publications in which I’ve been trying to lay this out below.)

Stevan Harnad

Harnad, S. (ed.) (1987) Categorical Perception: The Groundwork of Cognition. New York: Cambridge University Press.

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Harnad, S. (1990a) The Symbol Grounding Problem. Physica D 42: 335-346.

Harnad, S. (1990b) Against Computational Hermeneutics. (Invited commentary of Eric Dietrich’s Computationalism) Social Epistemology 4: 167-172.

Harnad, S. (1990c) Lost in the hermeneutic hall of mirrors. Invited Commentary on: Michael Dyer: Minds, Machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence 2: 321 - 327.

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on the Virtual Mind. Minds and Machines (in press)

Andrews, J., Livingston, K., Harnad, S. & Fischer, U. (1992) Learned Categorical Perception in Human Subjects: Implications for Symbol Grounding. Proceedings of Annual Meeting of Cognitive Science Society (submitted)

Harnad, S. Hanson, S.J. & Lubin, J. (1992) Learned Categorical Perception in Neural Nets: Implications for Symbol Grounding. Proceedings of Annual Meeting of Cognitive Science Society (submitted)

Harnad, S. (1993, in press) Icon, Category, Symbol: Essays on the Foundations and Fringes of Cognition. Cambridge University Press.

> Date: Tue, 31 Mar 1992 21:41:36 PST
> From: Pat Hayes
> Subject: Re: What is computation?
> To: Stevan Harnad
> Cc: chrisley@oxford.ac.uk
>
> Stevan-
>
> >SH: It seems to me that everything admits of a trivial computational
> >description.
>
> I have heard others say similar things, and Searle obviously believes
> something similar. Can you explain what you mean by this, and why you

> believe it? I cannot think of any sensible interpretation of this
> remark that makes it true. -- Pat

Pat, I think the trivial case is covered by Church's Thesis and Turing Equivalence. Consider a stone, just sitting there. It has one state, let's call it "0." Trivial computational description. Now consider the door, it has two states, open and shut; let's call one "0" and the other "1." Trivial computational description.

I think that's pretty standard, and has to do with the elementariness of the notion of computation, and hence that it can trivially capture, among other things, every static or simple dynamic description of a physical system. Nontrivial computation, on the other hand, is where Searle and I diverge. I think that if someone can define nontrivial computation in a principled way, it will separate computers from noncomputers (the way trivial computation does not).

Unfortunately, I do not have such a principled criterion for nontrivial computation except that (1) I think it will have a complexity-theoretic basis, perhaps related to NP-Completeness of the search for systematic rival interpretations of nontrivial symbol systems that differ radically from the standard interpretation (or its provable "duals"); and (2), even more vaguely, I feel the difference between trivial and nontrivial computation will be all-or-none rather than a matter of degree.

Earlier in this discussion it was pointed out that both Haugeland and Dennett (and now apparently McCarthy before them, and perhaps even Descartes -- see below) have also proposed a similar "cryptographer's constraint" on the nonarbitrariness of a systematic interpretation of a nontrivial symbol system (like natural language).

Stevan Harnad

> Date: Sun, 29 Mar 1992 21:40 EDT
> From: DDENNETT@PEARL.TUFTS.EDU
> Subject: Re: Cryptographer's Constraint
> To: harnad@Princeton.EDU
>
> I'm not sure when I FIRST discussed the cryptographers' constraint and I
> don't remember whether John McCarthy spoke of it before I did, but probably,
> since in True Believers (1981, reprinted in THE INTENTIONAL STANCE, 1987)
> I cite McCarthy 1979 when I mention it (p.29fn in TIS). By the way, the
> point can also be found in Descartes!!
> DAN DENNETT

Date: Wed, 1 Apr 92 09:09:47 EST From: "Stevan Harnad" To: hayes@sumex-aim.stanford.edu

> Date: Tue, 31 Mar 1992 23:22:44 PST
> From: Pat Hayes
>
> Stevan-
>
> OK, I now see what you mean: but what makes you think that calling the
> state of the stone '0' has anything to do with computation? A computer
> is a mechanism whose behavior is determined (in part) by the symbols

> stored in it. But the behavior of the stone and the door are not
> influenced in any way by the 'symbols' that this exercise in
> state-naming hypothesises. So they aren't computers.
>
> Perhaps I am reaching towards what you are calling nontrivial
> computation: but it might be less confusing to just call this
> computation, and call 'trivial computation' something else, might it
> not? What motivates this trivialisation of the computational idea?
>
> Pat

Pat, alas, what "trivializes" the computational idea is Goedel, Turing, Church, Post, von Neumann, and all the others who have come up with equivalent formulations of what computation is: It's just a very elementary, formal kind of thing, and its physical implementation is equally elementary. And by the way, the same problem arises with defining "symbols" (actually, "symbol-tokens," which are physical objects that are instances of an abstract "symbol-type"): For, until further notice, these too are merely objects that can be interpreted as if they meant something. Now the whole purpose of this exercise is to refute the quite natural conclusion that anything and everything can be interpreted as if it meant something, for that makes it look as if being a computer is just a matter of interpretation. Hence my attempt to invoke what others have apparently dubbed the "cryptographer's constraint" -- to pick out symbol systems whose systematic interpretation is unique and hard to come by (in a complexity-based sense), hence not arbitrary or merely dependent on the way we choose to look at them.

I also share your intuition (based on the programmable digital computer) that a computer is something that is mechanically influenced by its internal symbols (though we differ on two details -- I think it is only influenced by the SHAPE of those symbols, you think it's influenced by their MEANING [which I think would just put as back into the hermeneutic circle we're trying to break out of], and of course you think a conscious human implementation of a symbol system, as in Searle's Chinese Room, somehow does not qualify as an implementation, whereas I think it does). However, I recognize that, unlike in the case of formalizing the abstract notion of computation above, no one has yet succeeded in formalizing this intuition about physical implementation, at least not in such a way as to distinguish computers from noncomputers -- except as a matter of interpretation.

The "cryptographer's constraint" is my candidate for making this a matter of INTERPRETABILITY rather than interpretation, in the hope that this will get the interpreter out of the loop and let computers be computers intrinsically, rather than derivatively. However, your own work on defining "implementation" may turn out to give us a better way. To assess whether it succeeds, however, we're going to have to hear what your definition turns out to be! What you said above certainly won't do the trick.

One last point: As I've said before in this discussion, it is a mistake to conflate the question of what a computer is with the question of what a mind is. Even if we succeed in showing that computers are computers intrinsically, and not just as a matter of interpretation, there remains the independent problem of "intrinsic intentionality" (the fact that our thoughts are about what they are about intrinsically, and not just as a matter of interpretation by someone else). I, as you know, have recast this as the symbol grounding problem, and have concluded that, because of it, the implementation of a mind cannot possibly be merely the implementation of the "right" symbol system. There ARE other things under the sun besides computers, after all (indeed, confirming that is part of the goal of this exercise), and other processes besides (nontrivial) computation, and these will, I hypothesize, turn out to play an essential role in grounding MENTAL symbols, which are NOT sufficiently specified by their systematic interpretability alone: According to me, they must be grounded in the system's robotic interaction with the real world of objects that its symbols are "about," and this must likewise square systematically with the interpretation of its symbols. If you wish, this is a more rigorous "cryptographer's constraint," but this time a physical one rather than a merely a formal one. (Minds will accordingly turn out to be the TTT-scale class of "dedicated" computers, their "situated" "peripherals" and other analog structures and processes being essential substrates for their mental powers.)

Stevan Harnad

> Date: Wed, 1 Apr 92 09:59:32 -0500
> From: davism@turing.cs.nyu.edu (Martin Davis)
>
> Stevan,
>
> Thanks for keeping me posted on this debate. >
> I don't really want to take sides; however, there is technically no real
> problem in distinguishing "non-trivial" computers. They are "universal"
> if endowed with arbitrarily large memory.
>
> I've written two papers (long long ago) on the definition of universality
> for Turing machines. The first was in the McCarthy-Shannon collection
> "Automata Studies." The second was in the Proc. Amer. Math. Soc.
>
> If you want the exact references I'll be glad to forward them. But you
> may think this not relevant. Martin

Martin, I'm sure what you wrote will be relevant, so please do send me the reference. But can you also tell me whether you believe computers (in the real world) can be distinguished from noncomputers in any way that does not depend merely on how we choose to interpret their "states" (I think they can, Searle and others think they can't)? Do you think memory-size does it? Can we define "memory" interpretation- independently (to exclude, say, ocean tides from being computers)? And would your memory-size criterion mean that everything is a computer to some degree? Or that nothing is a computer, but some things are closer to being one than others? -- Cheers, Stevan

----

From: Ronald L Chrisley Date: Wed, 1 Apr 92 16:33:48 +0100

Stevan:

I think there is a large degree of agreement between us:

Date: Tue, 31 Mar 92 19:38:10 EST From: Stevan Harnad

> From: Ronald L Chrisley > Date: Wed, 25 Mar 92 16:13:22 GMT

SH: If I may interpolate some commentary: I agree about the physical grounding as picking out this machine running WORDSTAR as a privileged interpretation. I would add only two remarks.

(1) I think (though I can't prove it) that there is probably a complexity-based way of picking out the privileged interpretation of a system as a computer running a program (rather than other, more arbitrary interpretations) based on parsimony alone.

This may be true, but I think that "parsimony" here will probably have to make reference to causal relations.

(2) This discussion of what a computer is does not necessarily have any bearing on the question of what the mind is, or whether the brain is a computer. One could argue yes or no that computers/computation pick out a nonarbitrary kind. And one can independently argue yes or no that this has any substantive bearing on what kind of system can have a mind. (E.g., I happen to agree with Searle that a system will not have a mind merely because it implements the right computer program -- because, according to me, it must also be robotically grounded in the world -- but I disagree that there is no nonarbitrary sense in which some systems are computers and others are not. I.e., I agree with him about [intrinsic] semantics but not about syntax.)

I agree. I only mentioned the cognitivist's claim for perspective. Searle's claim that physics does not determine syntax is indeed distinct from his claim that syntax does not determine semantics. I'm very sympathetic with a grounded, embodied understanding of cognition. But that doesn't mean that I have to agree with Searle that the claim "mind is computation" is incoherent; it might just be wrong.

SH: It seems to me that everything admits of a trivial computational description.

I pretty much said the same when I said even a stone could admit of a computational description, and that such a notion of compoutation is unenlightening. But consider: perhaps the injustice in South Africa is something that does not even admit of a trivial computational description...

Only things with a certain kind of (not yet adequately specified) complexity admit of a nontrivial computational description (and those are computers). Now things that have minds will probably also admit of nontrivial computational descriptions, hence they too will be computers, but only in a trivial sense insofar as their MENTAL capacities are concerned, because they will not be ONLY computers, and their noncomputational robotic properties (e.g., transducers/actuators and other analog structures and processes) will turn out to be the critical ones for their mental powers; and those noncomputational properties will at the same time ground the semantics of the system's symbolic states.

This might be; but we should nevertheless resist Searle's following claim:

> > JS: If the question is : "Is the wall/brain INTRINSICALLY a digital
> > computer?" the answer is: NOTHING is intrisically a digital computer.
> > Please explain this point to your colleagues. they seem to think the
> > issue is universal realizability. Thus Chrisley's paper for example.

(BTW: was Searle assuming that the others had read/heard of my paper?!)

SH: I unfortunately can't explain this for Searle, because I happen to disagree with him on this point, although I do recognize that no one has yet come up with a satisfactory, principled way of distinguishing computers from noncomputers...

I agree.

SH: I don't think you'll be able to get computer scientists or physicists excited about the factor of "causality" in the abstract, but IMPLEMENTATION is certainly something they think about and have views on, because a program is just as an abstraction until and unless it's implemented (i.e., realized in a dynamical physical ["causal"] system -- a computer).

But Searle and Putnam have a point here: unless causality counts in determining what is and what is not an implementation, then just about anything can be seen as an implementation of anything else. So those interested in implementation will have to pay attention to causality.

But there's still not much room for a convergence of views there, because good "symbolic functionalists" hold that all the particulars of implementation are irrelevant -- i.e., that the same program can be implemented in countless radically different ways with nothing in common except that they are all implementations of the same computer program. Hence the right level to talk about is again the purely symbolic (copmputational) one.

But perhaps the point to be made is that there's a lot more involved in implementation than we previously realized. Symbolic functionalists knew that it placed some restriction on the physics; perhaps they just under-estimated how much.

I happen to disagree with these symbolic functionalists insofar as the mind is concerned, but not because I think there is something magic about the "causality" of implementation, but because I think a symbol system is just as ungrounded when it's implemented as when it's just scratches on static paper. The mere implementation of a program on a computer is the wrong kind of "causality" if a mind is what you're interested in implementing (or even if it's an airplane or a furnace). What's needed is the robotic (TTT) power to ground the interpretations of its internal symbols in the robot's interactions with the real world of objects, events and states of affairs that its symbols are interpretable as being "about" (TTT-indistinguishably from our own interactions with the world). (I list some of the publications in which I've been trying to lay this out below.)

Yes, I'm very sympathetic with your writings on this point. Even though the claim that everything realizes every Turing machine is false, that merely makes the claim "to have a mind is to implement TM No. xxx" coherent and false, not coherent and true. One still needs grounding.

But the reverse is also true. In a section of my paper ("Symbol Grounding is not sufficient"), I pointed out that one thing we can take home from Searle's paper is that without some appeal to causation, etc., in order to justify computational predicates, symbol grounding is mere behaviorism. We can agree with Searle on that and yet believe 1) that we *can* make the necessary appeals to causation in order to make sense of computational predicates (such appeals are implicit in our practice and theory); and 2) that symbol grounding, although not sufficient, is necessary for a computational understanding of mind.

Ronald L. Chrisley New College

> Date: Wed, 1 Apr 92 18:55:28 -0500
> From: davism@turing.cs.nyu.edu (Martin Davis)
> Subject: Re: What is a computation?
>
> Here are the references:
>
> “A Note on Universal Turing Machines,” {Automata Studies}, C.E.
> Shannon and J. McCarthy, editors, Annals of Mathematics Studies,
> Princeton University Press, 1956.
>
> “The Definition of Universal Turing Machine,” {Proceedings of the
> American Mathematical Society,} vol.8(1957), pp. 1125-1126.
>
> As for your argument with Searle (which I did try to avoid), my
> tendency is to place the issue in the context of the appropriate
> mathematical [idea] of "computer". I think it is a commonplace among
> philosophers that what appear to be purely empirical questions almost
> always really involve theoretical presuppositions.
>
> The two main contenders are finite automata and Turing machines. I
> suppose anything could be regarded as a finite automaton; I haven’t
> really thought about it. But most agree today (this wasn’t always the
> case) that it’s the TM model that’s appropriate. The counter-argument
> that real-world computers have finite memories is answered by noting
> that an analysis that defines a computer as having fixed memory size
> must say what kind of memory (ram? hard disk? floppies? tape?). In
> particular none of the theorems about finite automata have ever been
> applied to computers. If I remember (an increasingly dubious
> proposition) I discussed this in:
>
> “Computability,” {Proceedings of the Symposium on System
> Theory,} Brooklyn, N.Y. 1966, pp. 127-131.
>
> I would add (as I suggested in my previous message) that UNIVERSALITY
> is also generally tacitly presumed. This means that the computer can
> run programs embodying arbitrary algorithms.
>
> I think Searle would find it difficult to argue that a rock is a
> universal Turing machine.
>
> It is true that something may be a computer without it being readily

> recognized as such. This is for real. Microprocessors (which are
> universal computers) are part of many devices. Your telephone, your
> thermostat, certainly your VCR are all computers in this sense.
>
> But certainly not a rock!
>
> Here's another (closely related) less theoretical approach:
>
> Make a list of half a dozen simple computational tasks:
>
> E.g.
> 1. Given a positive integer, compute its square root to 5 decimal
> places.
> 2. Given two character strings, produce the string obtained by
> interleaving them, one character from each input at a time.
> 3. Given a positive integer, compute the sum of the positive integers
> less than or equal to the given integer;
>
> etc. etc.
>
> Then ask Searle to explain how to arrange matters so a stone will
> carry out these tasks.
>
> In other words, in order for the term "computer" to be justified, the
> object in question should be able to carry out ordinary computational
> tasks.
>
> Martin Davis

Martin,

Because I tend to agree with you in believing that a principled and interpretation-independent basis can be found for determining what is and is not a computer (and perhaps universality will be part of that basis), I'll leave it to other contributors to contest what you have suggested above. I do want to point out, however, that what we are trying to rule out here is arbitrary, gerrymandered interpretations of, say, the microstructure (and perhaps even the surface blemishes) of a stone according to which they COULD be mapped into the computations you describe. Of course the mapping itself, and the clever mind that formulated it, would be doing all the work, not the stone, but I think Searle would want to argue that it's no different with the "real" computer! The trick would be to show exactly why/how that rejoinder would be incorrect. It is for this reason that I have groped for a complexity-based (cryptographic?) criterion, according to which the gerrymandered interpretation of the stone could somehow be ruled out as too improbable to come by, either causally or conceptually, whereas the "natural" interpretation of the SPARC running WORDSTAR would not.

Stevan Harnad

PS Pat Hayes has furnished yet another independent source for this "cryptographer's constraint":

> Date: Tue, 31 Mar 1992 23:25:36 PST
> From: Pat Hayes
> Subject: Re: What is computation?
>
> PS: I recall McCarthy telling me the idea of the cryptographers
> constraint in 1969 when I first came to the USA (or maybe 1971, on the
> second trip). It didn't seem to be such an important matter then, of
> course.
>
> Pat Hayes

From: Ronald L Chrisley Date: Wed, 1 Apr 92 16:33:48 +0100

Stevan:

I think there is a large degree of agreement between us:

Date: Tue, 31 Mar 92 19:38:10 EST From: Stevan Harnad

> From: Ronald L Chrisley > Date: Wed, 25 Mar 92 16:13:22 GMT

SH: If I may interpolate some commentary: I agree about the physical grounding as picking out this machine running WORDSTAR as a privileged interpretation. I would add only two remarks.

(1) I think (though I can't prove it) that there is probably a complexity-based way of picking out the privileged interpretation of a system as a computer running a program (rather than other, more arbitrary interpretations) based on parsimony alone.

This may be true, but I think that "parsimony" here will probably have to make reference to causal relations.

(2) This discussion of what a computer is does not necessarily have any bearing on the question of what the mind is, or whether the brain is a computer. One could argue yes or no that computers/computation pick out a nonarbitrary kind. And one can independently argue yes or no that this has any substantive bearing on what kind of system can have a mind. (E.g., I happen to agree with Searle that a system will not have a mind merely because it implements the right computer program -- because, according to me, it must also be robotically grounded in the world -- but I disagree that there is no nonarbitrary sense in which some systems are computers and others are not. I.e., I agree with him about [intrinsic] semantics but not about syntax.)

I agree. I only mentioned the cognitivist's claim for perspective. Searle's claim that physics does not determine syntax is indeed distinct from his claim that syntax does not determine semantics. I'm very sympathetic with a grounded, embodied understanding of cognition. But that doesn't mean that I have to agree with Searle that the claim "mind is computation" is incoherent; it might just be wrong.

SH: It seems to me that everything admits of a trivial computational description.

I pretty much said the same when I said even a stone could admit of a computational description, and that such a notion of compoutation is unenlightening. But consider: perhaps the injustice in South Africa is something that does not even admit of a trivial computational description...

Only things with a certain kind of (not yet adequately specified) complexity admit of a nontrivial computational description (and those are computers). Now things that have minds will probably also admit of nontrivial computational descriptions, hence they too will be computers, but only in a trivial sense insofar as their MENTAL capacities are concerned, because they will not be ONLY computers, and their noncomputational robotic properties (e.g., transducers/actuators and other analog structures and processes) will turn out to be the critical ones for their mental powers; and those noncomputational properties will at the same time ground the semantics of the system's symbolic states.

This might be; but we should nevertheless resist Searle's following claim:

> > JS: If the question is : "Is the wall/brain INTRINSICALLY a digital
> > computer?" the answer is: NOTHING is intrisically a digital computer.
> > Please explain this point to your colleagues. they seem to think the
> > issue is universal realizability. Thus Chrisley's paper for example.

(BTW: was Searle assuming that the others had read/heard of my paper?!)

SH: I unfortunately can't explain this for Searle, because I happen to disagree with him on this point, although I do recognize that no one has yet come up with a satisfactory, principled way of distinguishing computers from noncomputers...

I agree.

SH: I don't think you'll be able to get computer scientists or physicists excited about the factor of "causality" in the abstract, but IMPLEMENTATION is certainly something they think about and have views on, because a program is just as an abstraction until and unless it's implemented (i.e., realized in a dynamical physical ["causal"] system -- a computer).

But Searle and Putnam have a point here: unless causality counts in determining what is and what is not an implementation, then just about anything can be seen as an implementation of anything else. So those interested in implementation will have to pay attention to causality.

But there's still not much room for a convergence of views there, because good "symbolic functionalists" hold that all the particulars of implementation are irrelevant -- i.e., that the same program can be implemented in countless radically different ways with nothing in common except that they are all implementations of the same computer program. Hence the right level to talk about is again the purely symbolic (copmputational) one.

But perhaps the point to be made is that there's a lot more involved in implementation than we previously realized. Symbolic functionalists knew that it placed some restriction on the physics; perhaps they just under-estimated how much.

I happen to disagree with these symbolic functionalists insofar as the mind is concerned, but not because I think there is something magic about the "causality" of implementation, but because I think a symbol system is just as ungrounded when it's implemented as when it's just scratches on static paper. The mere implementation of a program on a computer is the wrong kind of "causality" if a mind is what you're interested in implementing (or even if it's an airplane or a furnace). What's needed is the robotic (TTT) power to ground the interpretations of its internal symbols in the robot's interactions with the real world of objects, events and states of affairs that its symbols are interpretable as being "about" (TTT-indistinguishably from our own interactions with the world). (I list some of the publications in which I've been trying to lay this out below.)

Yes, I'm very sympathetic with your writings on this point. Even though the claim that everything realizes every Turing machine is false, that merely makes the claim "to have a mind is to implement TM No. xxx" coherent and false, not coherent and true. One still needs grounding.

But the reverse is also true. In a section of my paper ("Symbol Grounding is not sufficient"), I pointed out that one thing we can take home from Searle's paper is that without some appeal to causation, etc., in order to justify computational predicates, symbol grounding is mere behaviorism. We can agree with Searle on that and yet believe 1) that we *can* make the necessary appeals to causation in order to make sense of computational predicates (such appeals are implicit in our practice and theory); and 2) that symbol grounding, although not sufficient, is necessary for a computational understanding of mind.

Ronald L. Chrisley New College

Date: Thu, 2 Apr 1992 16:27:18 -0500 From: Drew McDermott Cc: hayes@sumex-aim.stanford.edu, searle@cogsci.Berkeley.EDU,

Here's my two-cents worth on the "everything is a computer" discussion.

From: "Stevan Harnad"

Pat, I think the trivial case is covered by Church's Thesis and Turing Equivalence. Consider a stone, just sitting there. It has one state, let's call it "0." Trivial computational description. Now consider the door, it has two states, open and shut; let's call one "0" and the other "1." Trivial computational description.

> From: Pat Hayes
>
> Stevan-
>
> OK, I now see what you mean: but what makes you think that calling the
> state of the stone '0' has anything to do with computation?

Unfortunately, I think this explanation by Stevan is not what Searle meant. Searle means to say that "computers are in the mind of the beholder." That is, if I take a system, and wish to view it as performing a computational sequence S, I can map the thermal-noise states (or any other convenient ways of partitioning its physical states) into computational states in a way that preserves the sequence. Putnam makes a similar claim in an appendix to, I think, "Representation and Reality." A long discussion about this has been going on in comp.ai.philosophy.

I agree with Stevan that Searle is wrong, and that computation is no more a matter of subjective interpretation than, say, metabolism is. However, I differ on where the problem arises:

[Stevan:]

Pat, alas, what "trivializes" the computational idea is Goedel, Turing, Church, Post, von Neumann, and all the others who have come up with equivalent formulations of what computation is: It's just a very elementary, formal kind of thing, and its physical implementation is equally elementary. And by the way, the same problem arises with defining "symbols" (actually, "symbol-tokens," which are physical objects that are instances of an abstract "symbol-type"): For, until further notice, these too are merely objects that can be interpreted as if they meant something. Now the whole purpose of this exercise is to refute the quite natural conclusion that anything and everything can be interpreted as if it meant something, for that makes it look as if being a computer is just a matter of interpretation. Hence my attempt to invoke what others have apparently dubbed the "cryptographer's constraint" -- to pick out symbol systems whose systematic interpretation is unique and hard to come by (in a complexity-based sense), hence not arbitrary or merely dependent on the way we choose to look at them.

I also share your intuition (based on the programmable digital computer) that a computer is something that is mechanically influenced by its internal symbols (though we differ on two details -- I think it is only influenced by the SHAPE of those symbols, you think it's influenced by their MEANING [which I think would just put as back into the hermeneutic circle we're trying to break out of], and of course you think a conscious human implementation of a symbol system, as in Searle's Chinese Room, somehow does not qualify as an implementation, whereas I think it does). However, I recognize that, unlike in the case of formalizing the abstract notion of computation above, no one has yet succeeded in formalizing this intuition about physical implementation, at least not in such a way as to distinguish computers from noncomputers -- except as a matter of interpretation.

The "cryptographer's constraint" is my candidate for making this a matter of INTERPRETABILITY rather than interpretation, in the hope that this will get the interpreter out of the loop and let computers be computers intrinsically, rather than derivatively. However, your own work on defining "implementation" may turn out to give us a better way.

I don't think it matters one little bit whether the symbols manipulated by a computer can be given any meaning at all. As I hope I've made clear before, the requirement that computers' manipulations have a meaning has been 'way overblown by philosopher types. The real reason why not every system can be interpreted as a computer is that the exercise of assigning interpretations to sequences of physical states of a system does not come near to verifying that the system is a computer. To verify that, you have to show that the states are generated in a lawlike way in response to future events (or possible events). It seems to me that for Searle to back up his claim that his wall can be viewed as a computer, he would have to demonstrate that it can be used to compute something, and of course he can't.

This point seems so obvious to me that I feel I must be missing something. Please enlighten me.

-- Drew

Date: Mon, 6 Apr 92 00:31:21 EDT Message-Id: <9204060431.AA00673@psycho>To: mcdermott-drew@CS.YALE.EDU

> Date: Thu, 2 Apr 1992 16:27:18 -0500
> From: Drew McDermott
>
> Searle means to say that "computers are in the mind of the beholder."
> That is, if I take a system, and wish to view it as performing a
> computational sequence S, I can map the thermal-noise states (or any
> other convenient ways of partitioning its physical states) into
> computational states in a way that preserves the sequence. Putnam makes
> a similar claim in an appendix to, I think, "Representation and
> Reality." A long discussion about this has been going on in
> comp.ai.philosophy.
>
> I agree with Stevan that Searle is wrong, and that computation is no
> more a matter of subjective interpretation than, say, metabolism is.
> However, I differ on where the problem arises...
>
> I don't think it matters one little bit whether the symbols manipulated
> by a computer can be given any meaning at all. As I hope I've made
> clear before, the requirement that computers' manipulations have a
> meaning has been 'way overblown by philosopher types. The real reason
> why not every system can be interpreted as a computer is that the
> exercise of assigning interpretations to sequences of physical states
> of a system does not come near to verifying that the system is a
> computer. To verify that, you have to show that the states are
> generated in a lawlike way in response to future events (or possible
> events). It seems to me that for Searle to back up his claim that his
> wall can be viewed as a computer, he would have to demonstrate that it
> can be used to compute something, and of course he can't.
>
> This point seems so obvious to me that I feel I must be missing
> something. Please enlighten me.
>
> -- Drew McDermott

Drew, I don't think anybody's very interested in uninterpretable formal systems (like Hesse's "Glass Bead Game"). Not just computational theory, but all of formal mathematics is concerned only with interpretable formal systems. What would they be otherwise? Just squiggles and squoggles we can say no more about (except that they follow arbitrary systematic rules like, "after a sguiggle and a squiggle comes a squaggle," etc.)? Now if THAT were all computation was, I would be agreeing with Searle!

It’s precisely the fact that it’s interpretable as amounting to MORE than just meaningless syntax that makes computation (and formal symbol systems in general) special, and of interest. And you yourself seem to be saying as much when you say "you have to show that the states are generated in a lawlike way in response to future events (or possible events)." For if this can be shown, then, among other things, it will also have been shown that they were interpretable. And, by the way, I don’t think that a computer playing, say, backgammon, is going to be shown to be a computer in virtue of "lawlike responses to future and possible events." It’s a computer because its states can be systematically interpreted as playing backgammon -- and a lot of other things (as suggested by those who have been stressing the criterion of universality in this discussion).

Now I really don’t think anything (even the human mind) can be coherently said to "respond" to future (or possible) events, whether in a lawlike or an unlawlike way (what is "lawlike," anyway, -- "interpretable as if governed by a law"?). So I can’t see how your proposed criteria help. But to answer your question about Searle: He didn’t say his wall could be USED to compute something, he said it could be DESCRIBED as if it were computing something. And you say as much in your own first paragraph.

Can you take a second pass at making your intutions explicit about this "lawlike performance" criterion, and how it separates computers from the rest? I think your criterion will have to be independent of the uses we may want to put the computer to, because making their computerhood depend on our uses sounds no better than making it depend on our interpretations.

Stevan Harnad

Date: Fri, 3 Apr 1992 16:31:05 PST From: Pat Hayes Subject: Re: What is computation? To: Stevan Harnad Cc: searle@cogsci.Berkeley.EDU, mcdermott@CS.YALE.EDU, hayes@cs.stanford.edu

Stevan,

I think that you (and others) are making a mistake in taking all the mathematical models of computation to be DEFINITIONS of computation. What makes it tempting to do so, I think, is the remarkable (and surprising) phenomenon of universality: that apparently any computer can be simulated on any other one, with enough resources. Trying to prove this led the theoretical folks in the forties to seek a definition, and it was tempting to choose some very simple device and say that THAT defined computation, since universality meant that this wasn’t any kind of restriction, it seemed, on what could (possibly) be computed. This enabled some good mathematics to be developed, but it was only a leap of faith, rather like the P/=NP hypothesis now: indeed it was actually called Church’s Thesis, if you recall. And as time has gone by it seems like it must be true, and one can kind of see why.

But to look in the literature and say that this means that computers are DEFINED to be, say, Turing machines, or any other kind of mathematical object, is just a philosophical mistake. You can’t run a Turing machine, for one thing, unless its engineered properly. (For example, the symbols on the tape would have to be in a form in which the processing box could read them, which rules out thermodynamic states of walls or rolls of toilet paper with pebbles on, and so forth.)

You might respond, well, what IS a computer, then? And my answer would be that this is essentially an empirical question. Clearly they are remarkable machines which have some properties unlike all other artifacts. What are the boundaries of the concept? Who knows, and why should I really care very much? For example, are neural net programs a form of computer, or are they something completely different? I would be inclined to say the former, but if someone wants to draw sharp lines excluding them, thats just a matter of terminology.

One point from your last message for clarification:

> a computer is something that is mechanically influenced by its
> internal symbols (though we differ on two details -- I think it is only
> influenced by the SHAPE of those symbols, you think it's influenced by
> their MEANING..

No, I don't think that the processor has access to anything other than the shape of the symbols (except when those symbols denote something internal to the machine itself, as when it is computing the length of a list: this point due to Brian Smith). I think we agree on this. But sometimes that suffices to cause the machine to act in a way that is systematically related to the symbol's meaning. All the machine has is some bitstring which is supposed to mean 'plus', but it really does perform addition.

For the record, I agree with you about the need for grounding of symbols to ultimately attach them to the world they purport to denote, but I also think that language enables us to extend this grounding to almost anything in the universe without actually seeing (feeling/hearing/etc) it, to the extent that the sensory basis of the glue is almost abstracted. One could imagine making a program which 'knew' a trmendous amount, could 'converse' well enough to pass the Turing Test in spades, etc., but be blind, deaf, etc.: a brain in a box. I think that its linguistic contact would suffice to say that its internal representations were meaningful, but you would require that it had some sensory contact. If we gave it eyes, you would say that all its beliefs then suddenly acquired meaning: its protests that it could remember the time when it was blind would be denied by you, since it would not have been nailed down sufficiently to the world then. Ah no, you would say to it: you only THOUGHT you knew anything then, in fact I KNOW you knew nothing. While I would have more humility.

best wishes

Pat Hayes

> Date: Fri, 3 Apr 1992 16:31:05 PST
> From: Pat Hayes
>
> You can't run a Turing machine, for one thing, unless its engineered
> properly. (For example, the symbols on the tape would have to be in a
> form in which the processing box could read them, which rules out
> thermodynamic states of walls or rolls of toilet paper with pebbles on,
> and so forth.)

>
> You might respond, well, what IS a computer, then? And my answer would
> be that this is essentially an empirical question. Clearly they are
> remarkable machines which have some properties unlike all other
> artifacts. What are the boundaries of the concept? Who knows, and why
> should I really care very much?

I agree it’s an empirical question, but it’s an empirical question we better be prepared to answer if there is to be any real substance to the two sides of the debate about whether or not the brain is (or is merely) a computer, or whether or not a computer can have a mind.

If Searle is right about the Chinese Room (and I am right about the Symbol Grounding Problem) AND there ARE things that are computers (implemented symbol-manipulating systems) as well as things that are NOT computers, then the former cannot have minds merely in virtue of implementing the right symbol system.

But if Searle is right about the "ungroundedness" of syntax too (I don’t happen to think he is), the foregoing alternatives are incoherent, because everything is a computer implementing any and every symbol system.

> No, I don’t think that the processor has access to anything other than
> the shape of the symbols (except when those symbols denote something
> internal to the machine itself, as when it is computing the length of a
> list: this point due to Brian Smith). I think we agree on this. But
> sometimes that suffices to cause the machine to act in a way that is
> systematically related to the symbol’s meaning. All the machine has is
> some bitstring which is supposed to mean ’plus’, but it really does
> perform addition.

I’m not sure what really performing addition is, but I do know what really meaning "The cat is on the mat" is. And I don’t think that when either an inert book or a dynamical TT-passing computer produces the string of symbols that is systematically interpretable as meaning "The cat is on the mat" (in relation to all the other symbols and their combinations) it really means "The cat is on the mat." And that is the symbol grounding problem. I do believe, however, that when a TTT-passing robot’s symbols are not only (1) systematically interpretable, but (2) those interpretations cohere systematically with all the robot’s verbal and sensorimotor interactions with the world of objects, events and states of affairs that the symbols are interpretable as being about, THEN when that robot produces the string of symbols that is systematically interpretable as "The cat is on the mat," it really means "The cat is on the mat."

> For the record, I agree with you about the need for grounding of
> symbols to ultimately attach them to the world they purport to denote,
> but I also think that language enables us to extend this grounding to
> almost anything in the universe without actually seeing
> (feeling/hearing/etc) it, to the extent that the sensory basis of the
> glue is almost abstracted. One could imagine making a program which
> ’knew’ a tremendous amount, could ’converse’ well enough to pass the
> Turing Test in spades, etc., but be blind, deaf, etc.: a brain in a
> box. I think that its linguistic contact would suffice to say that its

> internal representations were meaningful, but you would require that it
> had some sensory contact. If we gave it eyes, you would say that all
> its beliefs then suddenly acquired meaning: its protests that it could
> remember the time when it was blind would be denied by you, since it
> would not have been nailed down sufficiently to the world then. Ah no,
> you would say to it: you only THOUGHT you knew anything then, in fact I
> KNOW you knew nothing. While I would have more humility.
>
> best wishes Pat Hayes

Part of this is of course sci-fi, because we're not just imagining this de-afferented, de-efferented entity, but even imagining what capacities, if any, it would or could have left under those conditions. Let me say where I think the inferential error occurs. I can certainly imagine a conscious creature like myself losing its senses one by one and remaining conscious, but is that imagined path really traversable? Who knows what would be left of me if I were totally de-afferented and de-efferented. Note, though, that it would not suffice to pluck out my eye-balls, puncture my ears and peel off my skin to de-afferent me. You would have to remove all the analog pathways that are simply inward extensions of my senses. If you kept on peeling, deeper and deeper into the nervous system, removing all the primary and secondary sensory projections, you would soon find yourself close to the motor projections, and once you peeled those off too, you'd have nothing much left but the "vegetative" parts of the brain, controlling vital functions and arousal, plus a few very sparse and enigmatic sensory and sensorimotor "association" areas (but now with nothing left to associate) -- nor would what was left in any way resemble the requisite hardware for a computer (whatever that might be)!

Sure language is powerful, and once it's grounded, it can take you into abstractions remote from the senses; but I would challenge you to try to teach Helen Keller language if she had not been only deaf and dumb, but she had had no sensory or motor functions at all!

But never mind all that. I will remain agnostic about what the robot has to have inside it in order to have TTT power (although I suspect it resides primarily in that analog stuff we're imagining yanked out here), I insist only on the TTT-passing CAPACITY, not its necessarily its exercise. Mine is not a "causal" theory of grounding that says the word must "touch" its referent through some mystical baptismal "causal chain." The reason, I think, a person who is paralyzed and has lost his hearing, vision and touch might still have a mind is that the inner wherewithal for passing the TTT is still intact. But we know people can pass the TTT. A mystery candidate who can only pass the TT but not the TTT is suspect, precisely because of Searle's Argument and the Symbol Grounding Problem, for if it is just an implemented symbol system (i.e., a "computer" running a program), then there's nobody home in there.

The "need for grounding of symbols" is not merely "to ultimately attach them to the world they purport to denote," it is so that they denote the world on their own, rather than merely because we interpret them that way, as we do the symbols in a book.

Stevan Harnad

: Thu, 2 Apr 92 10:28:14 EST
> From: lammens@cs.Buffalo.EDU (Joe Lammens)
> Subject: symbol grounding
>
> Re: your article on "The Symbol Grounding Problem" in Physica D
> (preprint). If higher-order symbolic representations consist of symbol
> strings describing category membership relations, e.g. "An X is a Y
> that is Z", then who or what is doing the interpretation of these
> strings? They are just expressions in a formal language again, and I
> assume there is no grounding for the operators of that language like
> "is a" or "that is", whatever their actual representation? Even if
> there is, something still has to interpret the expressions, which seems
> to lead to a homunculus problem, or you'll have to define some sort of
> inferential mechanism that reasons over these strings. The latter seems
> to take us back to the realm of "traditional" AI completely, albeit
> with grounded constant symbols (or at least, some of them would be
> directly grounded). Is that what you had in mind? I don't see how in
> such a setup manipulation of symbols would be co-determined by the
> grounded meaning of the constant symbols, as you seem to require.
>
> Joe Lammens

Joe:

Your point would be valid if it were not for the fact that "Y" and "Z" in the above are assumed (recursively) to be either directly grounded or grounded indirectly in something that is ultimatly directly grounded. "X" inherits its grounding from Y and Z. E.g., if "horse" is directly grounded in a robot's capacity to identify (and discriminate and manipulate) horses on the basis of the sensorimotor interactions of the robot's transducers and effectors with horses, and "stripes" is likewise grounded, then "Zebra" in "A Zebra is a Horse with Stripes" inherits that grounding, and the proof of it is that the robot can now identify (etc.) a zebra upon its very first (sensorimotor) encounter with one. Such is the power of a grounded symbolic proposition.

To put it another way, the meanings of the symbols in a grounded symbol system must cohere systematically not only with (1) the interpretations we outsiders project on them (that's a standard symbol system), but also with (2) all of the robot's interactions with the world of objects, events and states of affairs that the symbols are interpretable as being about. No outsider or homunculus is needed to mediate this systematic coherence; it is mediated by the robot's own (TTT-scale) performance capacity, and in particular, of course, by whatever the internal structures and processes are that underlie that successful capacity. (According to my own particular grounding model, these would be analog projections connected to arbitrary symbols by neural nets that learn to extract the invariant features that make it possible for the robot to categorize correctly the objects of which they are the projections.)

A grounded symbol system is a dedicated symbol system, hence a hybrid one. In a pure symbol system, the "shape" of a symbol is arbitrary with respect to what it can be interpreted as standing for, and this arbitrary shape is operated upon on the basis of formal rules (syntax) governing the symbol manipulations. The only constraints on the manipulations are formal, syntactic ones. The remarkable thing about such pure symbol systems is that the symbols and symbol manipulations can be given a coherent systematic interpretation (semantics). Their short-coming, on the other hand (at least insofar as their suitability as models for cognition is concerned), is that the interpretations with which they so systematically cohere are nevertheless not IN the symbol system (any more than interpretations are in a book): They are projected onto them from the outside by us.

A grounded symbol system, by contrast, has a second set of constraints on it, over and above the syntactic ones above (indeed, this second set of constraints may be so overwhelming that it may not be useful to regard grounded symbol systems as symbol systems at all): The manipulation of both the directly grounded symbols and the indirectly grounded symbols (which are in turn grounded in them) is no longer contrained only by the arbitrary shapes of the symbols and the syntactic rules operating on those shapes; it is also constrained (or "co-determined," as you put it) by the NON-arbitrary shapes of the analog projections to which the ground-level symbols are physically connected by the category-invariance detectors (and, ultimately, to the objects those are the projections of). Indeed, because grounding is bottom-up, the non-arbitrary constraints are primary. ~X" is not free to enter into symbolic combinations except if the category relations the symbols describe square with the analog dictates of the ground-level symbols and their respective connections with the analog world of objects.

And the reason I say that such a dedicated symbol system may no longer even be usefully regarded as a symbol system at all can be illustrated if you try to imagine formal arithmetic -- Peano's Axioms, the formal rules of inference, and the full repertoire of symbols: "0", "1" "=" "+", etc. -- with all the elementary symbols "hard-wired" to the actual real-world quantities and operations that they are interpretable as referring to, with all symbol combinations rigidly constrained by those connections. This would of course not be formal arithmetic any more, but a "dedicated model." (I don't think it would be a good model for arithmetic COGNITION, by the way, because I don't think the elementary arithmetic symbols are directly grounded in this way; I'm just using it to illustrate the radical effects of nonarbitrary shape constraints on a formal system.)

So you see this is certainly not traditional AI. Nor is it homuncular. And what inferences it can make are hewing to more than one drummer -- the "higher" one of syntax and logic, but also the "lower" one of causal causal connections with the analog world of objects. And I do think that categorization is primary, rather than predication; to put it another way, predication and its interpretation is grounded in categorization. There is already categorization involved in "asserting" that an object is "Y." Conjunction may be an innate primitive, or it may be a primitive learned invariant. But once you can assert that this is a horse by reliably identifying it as "Horse" whenever you encounter it, and once you can do the same with "Stripes," then you are just a blank symbol away from identifying whatever has a conjunction of their invariants as "Zebra" (if that's the arbitrary symbol we choose to baptize it with).

Stevan Harnad

NOT POSTED
> Date: Mon, 06 Apr 92 11:01:12 ADT
> From: GOLDFARB%unb.ca@UNBMVS1.csd.unb.ca
>
> Stepa,
> I don't know what exactly Searle had in mind, but I also don't see
> anything interesting behind the idea of "computation". Every "object",
> including an "empty space" ("there is no space 'empty of field" --

> Einstein), might be said to perform many computations, depending
> on interactions with various other "objects", some of the computations
> are highly non-trivial.
>
> I think that "intelligent" computation is a more interesting idea to
> pursue: it is a degree to which a "system" is able to modify
> autonomously and irreversibly its INTERNAL states -- not just some
> auxiliary external objects, or symbols, as does the Turing machine --
> that have effect on all the related consequent computations.
>
> Cheers,
> Lev

Leva, I cannot post this to the list because as it stands it is immediately and trivially satisfied by countless actual computers running countless actual programs. I think you will have to follow the discussion a little longer to see what is at issue with this question of what is computation and what is a computer. Quick, vague, general criteria just won’t resolve things. -- Stepa

----

Date: Mon, 6 Apr 92 01:25:31 EST From: David Chalmers To: harnad@Princeton.EDU

I don’t think there’s a big problem here. Of course an answer to the question of whether "everything is a computer" depends on a criterion for when a computer, or a computation, is being physically implemented. But fairly straightforward criteria exist. While there is certainly room for debate about just what should be included or excluded, any reasonable criterion will put strong constraints on the physical form of an implementation: essentially, through the requirement that the state-transitional structure of the physical system mirror the formal state-transitional structure of the computation.

Start with finite state automata, which constitute the simplest formalism for talking about computation. An FSA is fixed by specification of a set of states S1,...,Sn, a set of inputs I1,...,Im, and a set of state-transition rules that map pairs to states. We can say that a given physical system implements the FSA if there is a mapping F from physical states of the system onto states of the FSA, and from inputs to the system onto inputs to the FSA, such that the state-transitional structure is correct: i.e. such that whenever the system is in physical state s and receives input i, it transits into a physical state s’ where maps to F(s’) according to the specification of the FSA.

This might look complex, but it’s very straightforward: the causal structure of the physical system must mirror the formal structure of the FSA, under an appropriate correspondence of states.

Some consequences:

(1) Any physical system will implement various FSAs -- as every physical system has *some* causal structure. e.g. the trivial one-state FSA will be implemented by any system. There’s no single canonical computation that a given object is implementing; a given object might implement various different FSAs, depending on the state correspondence that one makes. To that extent, computation is "interest-relative", but that’s a very weak degree of relativity: there’s certainly a fact of the matter about whether a given system is implementing a given FSA.

(2) Given a particular complex FSA -- e.g. one that a computationalist might claim is sufficient for cognition -- it will certainly not be the case that most objects implement it, as most objects will not have the requisite causal structure. There will be no mapping of physical states to FSA states such that state-transitional structure is reflected.

Putnam has argued in _Representation and Reality_ that any system implements any FSA, but that is because he construes the state-transition requirement on the physical system as a mere material conditional -- i.e. as if it were enough to find a mapping so that pairs are followed by the right s' on the occasions that they happen to come up in a given time interval; and if never comes up, then the conditional is satisfied vacuously. Of course the computationalist should construe the conditional as a strong one, with counterfactual force: i.e. whenever and however comes up, it must be followed by the right s'. Putnam's mappings fail to satisfy this condition -- if were to have come up another time, there's no guarantee that s' would have followed. There has been a long and interesting discussion of this topic on comp.ai.philosophy.

(3) Maybe someone will complain that by this definition, everything is performing some computation. But that's OK, and it doesn't make computation a useless concept. The computationalist claim is that cognition *supervenes* on computation, i.e. that there are certain computations such that any implementation of that computation will have certain cognitive properties. That's still a strong claim, unaffected by the fact that all kinds of relatively uninteresting computations are being performed all over the place.

To the person who says "doesn't this mean that digestion is a computation", the answer is yes and no. Yes, a given digestive process realizes a certain FSA structure; but this is not a very interesting or useful way to see it, because unlike cognition, digestion does not supervene on computation -- i.e. there will be other systems that realize the same FSA structure but that are not performing digestion. So: particular instances of digestion may be computations in a weak sense, but digestion as a type is not. It's only useful to take a computational view for properties that are invariant over the manner in which a computation is implemented. (Of course, Searle argues that cognition is not such a property, but that's a whole different can of worms.)

Finite state automata are a weak formalism, of course, and many if not most people will want to talk in terms of Turing machines instead. The extension is straightforward. We say that a physical system realizes a given Turing machine if we can map states of the system to states of the Turing-machine head, and separately map states of the system to symbols on each Turing-machine tape square (note that there will be a separate mapping for each square, and for the head, and also for the position of the head if we're to be complete), such that the state-transitional structure of the system mirrors the state-transitional structure of the Turing machine. For a Turing machine of any complexity, this will be a huge constraint on possible implementations.

So far, in talking about FSAs and Turing machines, we've really been talking about what it takes to implement a computation, rather than a computer. To be a computer presumably requires even stricter standards -- i.e., that the system be universal. But that is straightforward: we can simply require that the system implement a universal Turing machine, using the criteria above.

Personally I think that the notion of "computation" is more central to cognitive science than the notion of "computer". I don't see any interesting sense in which the human mind is a universal computer. It's true that we have the ability to consciously simulate any given algorithm, but that's certainly not a central cognitive property. Rather, the mind is performing a lot of interesting computations, upon which our cognitive properties supervene. So it’s probably most useful to regard cognitive processes as implementing a given non-universal Turing machine, or even an FSA, rather than a universal computer.

So, it seems to me that there are very straightforward grounds for judging that not everything is a computer, and that although it may be true that everything implements some computation, that’s not something that should worry anybody.

Dave Chalmers.

From: Stevan Harnad

David Chalmers wrote:

>dc> an answer to the question of whether "everything is a computer" depends
>dc> on a criterion for when a computer, or a computation, is being
>dc> physically implemented. But fairly straightforward criteria exist...
>dc> the causal structure of the physical system must mirror the formal
>dc> structure of the FSA, under an appropriate correspondence of states...
>dc>
>dc> Given a particular complex FSA -- e.g. one that a computationalist
>dc> might claim is sufficient for cognition -- it will certainly not be the
>dc> case that most objects implement it, as most objects will not have the
>dc> requisite causal structure...
>dc>
>dc> Finite state automata are a weak formalism, of course, and many if not
>dc> most people will want to talk in terms of Turing machines instead. The
>dc> extension is straightforward... For a Turing machine of any complexity,
>dc> this will be a huge constraint on possible implementations...
>dc>
>dc> To be a computer presumably requires even stricter standards -- i.e.,
>dc> that the system be universal. But that is straightforward: we can
>dc> simply require that the system implement a universal Turing machine,
>dc> using the criteria above...
>dc>
>dc> ...there are very straightforward grounds for judging that not
>dc> everything is a computer, and that although it may
>dc> be true that everything implements some computation, that’s not
>dc> something that should worry anybody.

I agree with Dave Chalmers’s criteria for determining what computation and computers are, but, as I suggested earlier, the question of whether or not COGNITION is computation is a second, independent one, and on this I completely disagree:

>dc> The computationalist claim is that cognition *supervenes* on
>dc> computation, i.e. that there are certain computations such that any
>dc> implementation of that computation will have certain cognitive
>dc> properties.
>dc>
>dc> To the person who says "doesn't this mean that digestion is a
>dc> computation", the answer is yes and no. Yes, a given digestive process
>dc> realizes a certain FSA structure; but this is not a very interesting or
>dc> useful way to see it, because unlike cognition, digestion does not
>dc> supervene on computation -- i.e. there will be other systems that
>dc> realize the same FSA structure but that are not performing digestion.
>dc>
>dc> Personally I think that the notion of "computation" is more central to
>dc> cognitive science than the notion of "computer". I don't see any
>dc> interesting sense in which the human mind is a universal computer...
>dc> Rather, the mind is performing a lot of interesting computations, upon
>dc> which our cognitive properties supervene. So it's probably most useful
>dc> to regard cognitive processes as implementing a given non-universal
>dc> Turing machine, or even an FSA, rather than a universal computer.

"Supervenience" covers a multitude of sins (mostly sins of omission). Whatever system turns out to be sufficient for having a mind, mental states will "supervene" on it. I don't feel as if I've said much of a mouthful there.

But it is a much more specific hypothesis that what the mind will "supervene" on is the right computations. We've agreed that what's special about computation is that there are many different ways to implement the same computations. So if a mind supervenes on (the right) computations because of their computational properties (rather than because of the physical details of any particular implementation of them), then it must supervene on ALL implementations of those computations. I think Searle's Chinese Room Argument has successfully pointed out that this will not be so, at least in the case of Searle's own implementation of the hypothetical Chinese-TT-passing computations -- except if we're willing to believe that his memorizing and executing a bunch of meaningless symbols is sufficient to cause a second mind to "supervene" on what's going on in his head -- something I, for one, would not be prepared to believe for a minute.

Because of certain similarities (similarities that on closer scrutiny turn out to be superficial), it was reasonable to have at first entertained the "computationalist" thesis that cognition might be a form of computation (after all, both thoughts and computations are put together out of strings of "symbols," governed by rules, semantically interpretable; both have "systematicity," etc.). But, because of the other-minds problem, there was always a systematic ambiguity about the standard Turing Test for testing whether a candidate system really had a mind.

We thought TT-passing was a good enough criterion, and no more or less exacting than the everyday criterion (indistinguishability from ourselves) that we apply in inferring that any other body than our own has a mind. But Searle showed this test was not exacting enough, because the TT could in principle be passed by computations that were systematically interpretable as a life-long correspondence with a pen pal who was understanding what we wrote to him, yet they could also be implemented without any understanding by Searle. So it turns out that we would have been over-interpreting the TT in this case (understandably, since the TT is predicated on the premise that to pass it is to generate and respond to symbols in a way that is systematically interpretable as -- and indistinguishable in any way from -- a life-long correspondence with a real person who really understands what we are writing). Such a test unfortunately trades on a critical ambiguity arising from the fact that since the TT itself was merely verbal -- only symbols in and symbols out -- there MIGHT have been only computations (symbol manipulations) in between input and output.

Well now that Searle has shown that that's not enough, and the Symbol Grounding Problem has suggested why not, and what might in fact turn out to be enough (namely, a system that passes the robotic upgrade of the TT, the Total Turing Test, able to discriminate, identify and manipulate the objects, events and states of affairs that it's symbols are systematically interpretable as being "about" in a way that is indistinguishable from the way we do), it's clear that the only way to resolve the ambiguity is to turn to abandon the TT for the TTT. But it is clear that in order to pass the TTT a system will have to do more than just compute (it must transduce, actuate, and probably do a lot of analog processing), and the mind, if any, will have to "supervene" on ALL of that -- not just the computations, which have already been shown not to be mindful! Moreover, whatever real computation a TTT-passer will be doing, if any, will be "dedicated" computation, constrained by the analog constraints it inherits from its sensorimotor grounding. And transducers, for example, are no more implementation-independent than digestion is. So not every implementation of merely their computational properties will be a transducer (or gastrointestinal tract) -- some will be mere computational simulations of transducers, "virtual transducers," and no mind (or digestion) will "supervene" on that.

Stevan Harnad

Date: Mon, 23 Mar 92 18:45:50 EST From: Eric Dietrich

Stevan:

Maybe it's the season: sap is rising, bugs are buzzing, and trees are budding -- but it seems to me that some progress has been made on the question of computers, semantics, and intentionality. (BTW: thank you for bouncing to me your discussion with Searle. I enjoyed it.)

I agree with Searle on two points. First, nothing is intrinsically a computer. And second, the big problem is not universal realizability.

Furthermore, I agree with you that computation and implementation are not the same thing, and that nontrivial symbol systems will not have arbitrary duals because they have a certain complex systematicity.

But, ... 1. Nothing is intrinsically a computer because nothing is intrinsically anything. It's interpretation all the way down, as it were.

2. Therefore, it's lack of imagination that prevents us from swapping interpretations in general in English, arithmetic, and Lisp. This lack of imagination is, though, is part of our epistemic boundedness. We are not stupid, just finite. To keep things coherent, and to swap all the meanings in English is something that we cannot do. Perhaps no intelligent creature could do this because creatures vastly more intelligent than we would have that much more science -- explanations and semantics -- to juggle when trying to invent and swap duals.

3. Still, we arrive at the same point: a wall is only an implementation of a trivial turing machine or computation. . . .

But, ... How can we arrive at the same point if I believe that computers are NOT formal symbol manipulators while you and Searle believe that they are? Because computation is an observer relative feature precisely *because* semantics is. In other words, you can interpret your wall, there just isn't much reason to do so. Planets can be viewed as representing and computing their orbits, but there isn't much reason to do so. Why? Because it involves too much "paper work" for us. Other intelligent entities might prefer to attribute/see such computations to the planets.

For me, computation, systematicity, and semantics are matters of degree. Machines, computation, and meaning are in the eye of the beholder, or more precisely, the explainer.

What recommends this view? Does it give us exactly the same conclusions as your view? No, it is not the same. Interpretationalism provides a different set of problems that must be solved in order to build an intelligent artifact, problems that are prima facie tractable. For example, on the interpretationalist view, you don't have to solve the problem of original intentionality (or, what is the same, the problem provided by the Chinese Room); nor do you have to solve the symbol grounding problem (though you do have to figure out how perception and categorization works). You can instead spend your time searching for the algorithms (equivalently, the architectures) responsible for our intelligence -- architectures for plasticity, creativity and the like.

More deeply, it allows us the explanatory freedom to handle the computational surprises that are no doubt in our future. In my opinion, the semantical view espoused by you and Searle is too rigid to do that.

And finally, interpretationalism holds out the promise that cognitive science will integrate (integrate, NOT reduce) smoothly with our other sciences. If intentionality is a real property of minds, then minds become radically different from rocks. So different that I for one despair of ever explaining them at all. (Where, for example, do minds show up phylogenetically speaking? And why there and not somewhere else? These are questions YOU must answer. I don't have to.)

We don't need to preserve psychology as an independent discipline by giving it phenomena to explain that don't exist anywhere else in nature. Rather, we can preserve psychology because it furthers our understanding in a way that we would miss if we stopped doing it.

Sincerely,

Eric

"INTERPRETATIONALISM" AND ITS COSTS

Eric Dietrich wrote:

> ed> I agree with Searle [that] nothing is intrinsically a computer [and
> ed> that] the big problem is not universal realizability... I agree with
> ed> you that computation and implementation are not the same thing, and
> ed> that nontrivial symbol systems will not have arbitrary duals because

> ed> they have a certain complex systematicity... But, ... Nothing is
> ed> intrinsically a computer because nothing is intrinsically anything.
> ed> It’s interpretation all the way down, as it were.

A view according to which particles have mass and spin ond obey Newton’s laws only as a matter of interpretation is undesirable not only because it makes physics appear much more subjective and impressionistic than necessary, but because it blurs a perfectly good and informative distinction between the general theory-ladenness of all scientific inferences and the special interpretation-dependence of the symbols in a computer program (or a computer implementation of it). It is the latter that is at issue here. There is, after all, a difference between my "interpreting" a real plane as flying and my interpreting a computer simulation of a plane as flying.

> ed> ...it’s lack of imagination that prevents us from swapping
> ed> interpretations in general in English, arithmetic, and Lisp. This lack
> ed> of imagination, though, is part of our epistemic boundedness. We are
> ed> not stupid, just finite. To keep things coherent, and to swap all the
> ed> meanings in English is something that we cannot do. Perhaps no
> ed> intelligent creature could do this because creatures vastly more
> ed> intelligent than we would have that much more science -- explanations
> ed> and semantics -- to juggle when trying to invent and swap duals.

I don’t know any reasons or evidence for believing that it is lack of imagination that prevents us from being able to come up with coherent interpretations for arbitrarily swapped symbols. NP-completeness sounds like a good enough reason all on its own.

> ed> Still, we arrive at the same point: a wall is only an implementation of
> ed> a trivial turing machine or computation. But, ... How can we arrive at
> ed> the same point if I believe that computers are NOT formal symbol
> ed> manipulators while you and Searle believe that they are? Because
> ed> computation is an observer relative feature precisely *because*
> ed> semantics is. In other words, you can interpret your wall, there just
> ed> isn’t much reason to do so. Planets can be viewed as representing and
> ed> computing their orbits, but there isn’t much reason to do so. Why?
> ed> Because it involves too much "paper work" for us. Other intelligent
> ed> entities might prefer to attribute/see such computations to the
> ed> planets.

I think the reason planets don’t compute their orbits has nothing to do with paperwork; it is because planets are not computing anything. They are describable as computing, and the computation is implementable as a computer simulation of planetary motion (to an approximation), but that’s just because of the power of formal computation to approximate (symbolically) any physical structure or process at all (this is a variant of Church’s Thesis).

Allowing oneself to be drawn into the hermeneutic hall of mirrors (and leaving the virtual/real distinction at the door) can lead to illusory after-effects even when one goes back into the real world. For not only does one forget, while in the hall of mirrors, that the fact that computations are interpretable as planetary motions does not make them real planetary motions, but even when one re-enters the real world one forgets that the fact that planets are describable as computing does not mean they are really computing!

> ed> For me, computation, systematicity, and semantics are matters of
> ed> degree. Machines, computation, and meaning are in the eye of the
> ed> beholder, or more precisely, the explainer.

For me what distinguishes real planetary motion from a computer simulation of it is definitely NOT a matter of degree. Ditto for meaning and mind.

> ed> What recommends this view? Does it give us exactly the same conclusions
> ed> as your view? No, it is not the same. Interpretationalism provides a
> ed> different set of problems that must be solved in order to build an
> ed> intelligent artifact, problems that are prima facie tractable. For
> ed> example, on the interpretationalist view, you don't have to solve the
> ed> problem of original intentionality (or, what is the same, the problem
> ed> provided by the Chinese Room); nor do you have to solve the symbol
> ed> grounding problem (though you do have to figure out how perception and
> ed> categorization works). You can instead spend your time searching for
> ed> the algorithms (equivalently, the architectures) responsible for our
> ed> intelligence -- architectures for plasticity, creativity and the like.

I adopt the simple intermediate position that if the meanings of whatever symbols and computations are actually going on inside a robot are grounded (TTT-indistinguishably) in the robot's sensorimotor interactions (with the real world of objects that its symbols are systematically interpretable as being about), then there are no (solvable) problems left to solve, and the particular branch of reverse bioengineering that is "cognitive science" will have done its work, fully integrably with the rest of pure and applied science.

Of course, as with the computational modelling of planetry motion, a great deal can be found out (empirically and analytically) about how to get a robot to pass the TTT through simulations alone, but the simulation itself is not the TTT and the simulated robot does not have a mind. Alas, "interpretationalism" seems to lose this distinction.

> ed> interpretationalism holds out the promise that cognitive science will
> ed> integrate (integrate, NOT reduce) smoothly with our other sciences. If
> ed> intentionality is a real property of minds, then minds become radically
> ed> different from rocks. So different that I for one despair of ever
> ed> explaining them at all. (Where, for example, do minds show up
> ed> phylogenetically speaking? And why there and not somewhere else? These
> ed> are questions YOU must answer. I don't have to.)

Not at all. The do-able, empirical part of mind-modelling is TTT-modelling, and that can in principle (though not in practice) be accomplished for all species without ever having to answer the (unanswerable) question of where mind starts and who/what does/doesn't have a mind (apart from oneself). "Interpretationalism" can't answer the question either, but it disposes of it at the very high price of supposing (1) that everything has a mind to some degree and (2) that the (real/virtual) difference between having any physical property P and merely being systematically interpretable as having property P is no difference at all -- at the price, in other words, of simultaneously begging the question (2) and answering it by fiat (1)!

Stevan Harnad

Date: Tue, 7 Apr 92 23:56:53 PDT From: sereno@cogsci.UCSD.EDU (Marty Sereno) To: harnad@Princeton.EDU Subject: Cells, Computers, and Minds

hi stevan

I have patiently read the many posts on the symbol-grounding problem with interest for several years now. Many of the comments have floundered around trying to find a clear definition of what it takes to make a symbol-using system "really" understand something. They tend to get tied up with various human artifacts, and it can be extremely difficult to sort out the various sources of meaning-grounding. We can avoid some of these problems by considering cells, which have the distinction of being the first grounded symbol-using system--and one whose grounding does not depend on any human artifact, or on humans at all, for that matter.

The use of symbols strings in cells is well documented and rather different than the use of symbol strings in human-designed computers. The plan is to compare a computer to a cell, and then argue that human symbol use looks more like that in cells.

The basic difference can be quite simply stated. Computers consist of some kind of device that can read code strings and then write code strings in a systematic, programmable way (with due respect to what has been written on this topic). Reading and writing code is to perform some kind of binary-like classification of symbol tokens (e.g., reading 4.8 volts to be the same as 5 volts). Computer designers have found numerous ways to relate these written and read code strings to real world tasks (e.g., A/D and D/A convertors, operators who understand human and computer languages).

A cell reads code strings as well. Each living cell contains somewhere between 1 and 200 megabytes of code. Messenger RNA sequences transcribed from this permanent store are recognized by the cell during the process of protein "translation" to contain codons each containing 3 nucleotides. Each nucleotide can each be described as having two features: "long/short" (A and G [purines] vs. C and T [pyrimidines]) and "2/3 bonds" (A and T vs. G and C). The key point is that there are no other examples of *naturally-occurring* systems that use long code-strings like these that are conceivable *without* protein translation or human thought (this disqualifies the immune system and mathematical notation as independent naturally-occurring, self-maintaining systems, for me at least).

But the way cells put these recognized symbols to work is remarkably different than with computers. Instead of reading code for the purpose of *operating on other code*, cells use the code to make proteins (esp. enzymes), which they then use to maintain a metabolism. Proteins are constructed by simply bonding amino acids into an (initially) 1-D chain that is parallel to the recognized codons (words) in the messenger RNA chain. Amino acids have none of the characteristics of nucleotide symbol segment chains. Objective characteristics of (molecular) symbol segment chains for me are: similar 3-D structure despite 1-D sequence differences; small number of binary-like features for each segment; their use as a 1-D chain in which small groups of segments are taken to stand for a sequence of other, possibly non-symbolic things.

Proteins are extremely complex molecules, each containing thousands of atoms in a precise 3-D arrangement. The DNA sequences in the genome, however, constitute only a trivial portion of what would be required to explicitly specify the 3-D structure of a protein; a single gene typically contains only a few hundred bytes of information. This information goes such a long way because it depends for its interpretation on the existence of elaborate geometrical constraints due to covalent chemical bonding, weak electronic interactions, the hydrophobic effect, the structural details of the 20 amino acids, and so on--a large set of 'hard-wired' effects that the cell harnesses, but cannot change. Once the amino acid chain has been synthesized, its self-assembly (folding) is directed entirely by these prebiotic, non-symbolic chemical constraints.

Certain aspects of the architecture of cellular metabolism is much like a production system. The enzymes ("productions") of metabolism operate on their substrates ("objects") in a cytoplasm ("working memory"), which requires that they have a great deal of specificity to avoid inappropriate interactions. As in some kinds of production systems, enzymes can operate on other enzymes as substrates. The key difference is that the code in the cellular system is used strictly to make the enzyme "productions"; once they are made, they fold up and operate primarily in a non-symbolic milieu and on non-symbolic things in the cytoplasm (this not exclusively the case; some proteins do in fact control which part of the code is read).

No one in their right mind would want to make a computer more like a cell for most of things that computers are currently used for. It is much to hard too make arbitrary local changes in a cell's metabolism; and evolution takes a miserably long time and involves large populations. Molecular biologists, however, might, conversely, like to engineer a cell into a computer by using overzealous error-correcting polymerases to write DNA code. Code manipulations are not very fast and would probably have to be rather local in cells, but it would be easy to get billions or trillions of copies of a bacterial "program" in a short time.

I suggest that we might take a cue from how cellular symbols are grounded in thinking about how human symbols are grounded. Following the cellular architecture, we might conjecture that the main use of symbol strings for humans--in particular, external speech symbol strings--is to construct an internal "mental metabolism". Small groups of speech sounds are first internalized in auditory cortical areas, and then small groups of them are recognized and taken to stand for other non-symbolic internal patterns--e.g., visual category patterns in higher cortical visual areas. Perhaps, human language involves relying on pre-linguistic constraints on how sequentially activated and "bound together" visual category activity patterns interact in higher primate visual cortical areas. We could think of language as a kind of code-directed scene comprehension that relies on implicit harnessing of pre-existing constraints in a way analogous to the use of a complex chemistry by cellular code strings. There is a similar compactness to the code (a few hundred bytes of information specifies an enzyme and the complex meaning of a discourse in the mind of the listener). It is amazing to consider that the genetic code for an entire living, reproducing, self-maintaining E. coli bacterium takes up less space than the code for a decent word processor.

I would argue that a human-like symbol-using system depends on harnessing complex dynamical constraints in a non-symbolic world, just as cellular symbol systems depend on complex chemistry for their grounding. It is not likely to be easy to construct such a "chemistry" in an artificial machine. Real chemistry is extremely complex and the specification of protein structure relies on many intricate details of this complexity; it is not currently possible to predict the 3-D structure of a protein given only the amino acid sequence. The "chemistry" of interacting patterns in human neural networks is undoubtedly even more complex. But there may be no other way to make a grounded symbol-using system.

For a longer exposition of these ideas, see:

Sereno, M.I. (1991) Four analogies between biological and cultural/linguistic evolution. Journal of Theoretical Biology 151:467-507.

Sereno, M.I. (1991) Language and the primate brain. Proceedings, Thirteenth Annual Conference of the Cognitive Science Society, Lawrence Erlbaum Assoc., pp. 79-84.

Though my note is a little long, please print it out before singling out particular sentences for ridicule or praise...

marty

Date: Fri, 17 Apr 92 17:07:20 EDT From: "Stevan Harnad"

ON SYMBOL SYSTEMS: DEDICATED, GROUNDED AND CELLULAR

Marty Sereno (sereno@cogsci.UCSD.EDU) wrote:

ms> cells... have the distinction of being the first grounded symbol-using ms> system--and one whose grounding does not depend on any human artifact, ms> or on humans at all, for that matter... The use of symbols strings in ms> cells is well documented and rather different [from] the use of symbol ms> strings in human-designed computers... But the way cells put these ms> recognized symbols to work is remarkably different... Instead of ms> reading code for the purpose of *operating on other code*, cells use ms> the code to make proteins (esp. enzymes), which they then use to ms> maintain a metabolism... The key difference is that the code in the ms> cellular system is used strictly to make the enzyme "productions"; once ms> they are made, they fold up and operate primarily in a non-symbolic ms> milieu and on non-symbolic things in the cytoplasm... ms> ms> I would argue that a human-like symbol-using system depends on ms> harnessing complex dynamical constraints in a non-symbolic world, just ms> as cellular symbol systems depend on complex chemistry for their ms> grounding. It is not likely to be easy to construct such a "chemistry" ms> in an artificial machine... But there may be no other way to make a ms> grounded symbol-using system.

A cell seems to be like a dedicated computer. A dedicated computer is one for which the interpretations of some or all of its symbols are "fixed" by the fact that it is hard-wired to its input and output. In this sense, a dedicated chess-playing computer -- one whose inputs and outputs are pysically connected only to a real chess board and chess-men -- is a grounded symbol system (considered as a whole). Of course, a dedicated chess-playing computer, even though it is grounded, is still just a toy system, and toy systems are underdetermined in more ways than one. To ground symbol meanings in such a way as to make them completely independent of our interpretations (or at least no more nor less indeterminate than they are), a symbol system must be not only grounded but a grounded TTT-scale robot, with performance capacity indistinguishable from our own.

In a pure symbol system, the "shapes" of the symbols are arbitrary in relation to what they can be interpreted as meaning; in a dedicated or grounded symbol system, they are not. A cell seems to be more than just a dedicated computer, however, for mere dedicated computers still have sizeable purely computational components whose function is implementation-independent, hence they can be "swapped" for radically different physical systems that perform the same computations. In a dedicated chess-playing computer it is clear that a radically different symbol-manipulator could be hard-wired to the same input and output and would perform equivalent computations. It is not clear whether there are any implementation-independent components that could be swapped for radically different ones in a cell. This may either be a feature of the "depth" of the grounding, or, more likely, an indication that a cell is not really that much like a computer, even a dedicated one. The protein-coding mechanisms may be biochemical modules rather than formal symbols in any significant sense.

There’s certainly one sense, however, in which cells and cellular processes are not merely materials for analogies in this discussion, because for at least one TTT-passing system (ourselves) they happen to generate a real implementation! Now, although I am not a "symbolic" functionalist (i.e., I don’t believe that mental processes are implementation-independent in the same way that software is implementation-independent), I am still enough of a ("robotic") functionalist to believe that there may be more than one way to implement a mind, perhaps ways that are radically different from the cellular implementation. As long as they have TTT-indistinguishable performance capacity in the real world, I would have no nonarbitrary grounds for denying such robots had minds.

ms> I suggest that we might take a cue from how cellular symbols are ms> grounded in thinking about how human symbols are grounded. Following ms> the cellular architecture, we might conjecture that the main use of ms> symbol strings for humans--in particular, external speech symbol ms> strings--is to construct an internal "mental metabolism". Small groups ms> of speech sounds are first internalized in auditory cortical areas, and ms> then small groups of them are recognized and taken to stand for other ms> non-symbolic internal patterns--e.g., visual category patterns in ms> higher cortical visual areas. Perhaps, human language involves relying ms> on pre-linguistic constraints on how sequentially activated and "bound ms> together" visual category activity patterns interact in higher primate ms> visual cortical areas. We could think of language as a kind of ms> code-directed scene comprehension that relies on implicit harnessing of ms> pre-existing constraints in a way analogous to the use of a complex ms> chemistry by cellular code strings.

This analogy is a bit vague, but I would certainly be sympathetic to (and have indeed advocated) the kind of sensory grounding it seems to point toward.

Stevan Harnad

Date: Fri, 17 Apr 92 17:48:35 EDT From: "Stevan Harnad"

> Date: Mon, 6 Apr 92 20:02 GMT
> From: UBZZ011@cu.bbk.ac.uk Todd Moody
> To: HARNAD <@nsfnet-relay.ac.uk:HARNAD@PRINCETON.edu>
>
> Another way to ask the question at hand is to ask whether, given some

> alien object that appeared to be undergoing complex changes in its
> discrete state configurations, is it possible to tell by inspection
> whether it is doing computation? (alien means we don't know the
> "intended interpretation," if there is one, of the states) This
> question is rather strongly analogous to a question about language:
> Given some arbitrary complex performance (dolphin noise, for example),
> is it possible to determine whether it is a linguistic performance
> without also being able to translate at least substantial portions of
> it?
>
> In both cases, I don't see how the questions can be answered other than
> by working from considerations of *parsimony under interpretation*.
> That is, in the case of dolphin noise, you just have to make some
> guesses about dolphin interests and then work on possible
> interpretation/translations. When you reach the point that the simplest
> interpretation of the noise is that it means XYZ, then you have a
> strong case that the noise is language. In the case of the alien
> thing-that-might-be-a-computer, the trick is to describe it as
> following a sequence of instructions (or computing a function) such
> that this description is simpler than a purely causal description of
> its state changes.
>
> A description of an object as a computer is more *compressed* (simpler)
> than the description of it as an arbitrary causal system.
>
> Thus, it is parsimony under interpretation that rules out Searle's
> wall. This is not interpretation-independent, but I think it is as
> good as it gets.
>
> Todd Moody (tmoody@sju.edu)

Todd, I agree with this strategy for judging whether or not something is computing (it is like the complexity-based criterion I proposed, and the "cryptographic criterion" Dennett, Haugeland, McCarthy and perhaps Descartes proposed), but it won't do for deciding whether the interpretation is intrinsic or derived. For that, you need more than interpretability (since it already presupposes interpretability). My candidate is grounding in (TTT-scale) robotic interactions with the world of objects the symbols are interpretable as being about.

Stevan Harnad

From: Jeff Dalton Date: Mon, 6 Apr 92 18:54:27 BST

Steven Harnad writes:

> that what we are
> trying to rule out here is arbitrary, gerrymandered interpretations of,
> say, the microstructure (and perhaps even the surface blemishes) of a

> stone according to which they COULD be mapped into the computations you
> describe. Of course the mapping itself, and the clever mind that
> formulated it, would be doing all the work, not the stone, but I think
> Searle would want to argue that it's no different with the "real"
> computer! The trick would be to show exactly why/how that rejoinder
> would be incorrect. It is for this reason that I have groped for a
> complexity-based (cryptographic?) criterion, according to which the
> gerrymandered interpretation of the stone could somehow be ruled out as
> too improbable to come by, either causally or conceptually, whereas the
> "natural" interpretation of the SPARC running WORDSTAR would not.

One potential problem with the complexity constraint is that the interpretations are expressed in a particular language (let us say). An interpretation that is more complex in one language might be simpler in another. Putnam makes a similar point about his "cats are cherries" example, that which interpretation is the weird one switches depending on whether you're expressing the interpretation in the language where "cats" means cats or the one in which it means cherries.

As a metaphor for this, consider random dot stereograms as an encoding technique (something suggested to me by Richard Tobin). Someone mails you a picture that consists of (random) dots. Is it a picture of the Eiffel Tower, or a Big Mac? Well, they mail you another picture of random dots and, viewed together with the first, you see a picture of the Eiffel Tower. But they could just as well have mailed you a different second picture that, together with the first, gave a Big Mac.

Moreover, it is not true in general that the simpler interpretation is always the right one. Someone who is encoding something can arrange for there to be a simple interpretation that is incorrect. I suppose an example might be where the encrypted form can be decrypted to an English text, but the actual message can only be found buy taking the (English) words that appear after every third word that contains an "a".

-- jeff

Date: Sat, 18 Apr 92 12:58:51 EDT From: "Stevan Harnad"

COMPLEXITY, PARSIMONY and CRYPTOLOGY

Jeff Dalton wrote:

>jd> Stevan Harnad wrote:

>
>sh> what we are trying to rule out here is arbitrary, gerrymandered
>
>sh> interpretations of, say, the microstructure (and perhaps even the
>
>sh> surface blemishes) of a stone according to which they COULD be mapped
>
>sh> into the computations you describe. Of course the mapping itself, and
>
>sh> the clever mind that formulated it, would be doing all the work, not
>

>sh> the stone, but I think Searle would want to argue that it’s no
>
>sh> different with the "real" computer! The trick would be to show exactly
>
>sh> why/how that rejoinder would be incorrect. It is for this reason that I
>
>sh> have groped for a complexity-based (cryptographic?) criterion,
>
>sh> according to which the gerrymandered interpretation of the stone could
>
>sh> somehow be ruled out as too improbable to come by, either causally or
>
>sh> conceptually, whereas the "natural" interpretation of the SPARC running
>
>sh> WORDSTAR would not.

>jd> One potential problem with the complexity constraint is that the
>jd> interpretations are expressed in a particular language (let us say).
>jd> An interpretation that is more complex in one language might be
>jd> simpler in another. Putnam makes a similar point about his "cats
>jd> are cherries" example, that which interpretation is the weird one
>jd> switches depending on whether you’re expressing the interpretation
>jd> in the language where "cats" means cats or the one in which it
>jd> means cherries.

As I understand the Chaitin/Kolmogorov complexity-based criterion for parsimony and randomness (Chaitin 1975; Rabin 1977), an algorithm (a string of bits) is nonrandom and parsimonious to the degree that the number of bits in it is smaller than the number of bits in the "random" string (which is usually infinitely long) that it can be used to generate. The measure of parsimony is the relative size of the short ("theory") and long ("data") bit string. It is stressed that language and notational variations may alter the length of the algorithm by a few bits, but that all variants would still be an order of magnitude smaller than the data string (and therein lies the real parsimony).

Now I realize that the C/K criterion is only a thesis, but I think it conveys the intuition that I too would have: that the relative ease with which some things can be expressed in English rather than French (or FORTRAN rather than ALGOL) is trivial relative to the fact that they can be expressed at all, either way.

Two qualifications, however:

(1) The C/K criterion applies to algorithms as uninterpreted strings of bits that "generate" much longer uninterpreted strings of bits. The short and long strings are interpretable, respectively, as theory and data, but -- as usual in formal symbol systems -- the interpretation is external to the system; the counting applies only to the bits. So although I don’t think it is circular or irrelevant to invoke the C/K analogy as an argument for discounting linguistic and notational differences, doing so does not go entirely to the heart of the matter of the parsimony of an INTERPRETATION (as opposed to an uninterpreted algorithm).

(2) Another potential objection is more easily handled, however, and again without any circularity (just some recursiveness): When one is assessing the relative complexity of an algorithm string and the (much longer) data string for which it is an algorithm, the potential differences among the languages in which one formulates the algorithm (and the data) clearly cannot include potential gerrymandered languages whose interpretation itself requires an algorithm of the same order of magnitude as the data string! That's precisely what this complexity-based/cryptographic criterion is invoked to rule out!

>jd> As a metaphor for this, consider random dot stereograms as an encoding
>jd> technique (something suggested to me by Richard Tobin). Someone mails
>jd> you a picture that consists of (random) dots. Is it a picture of the
>jd> Eiffel Tower, or a Big Mac? Well, they mail you another picture of
>jd> random dots and, viewed together with the first, you see a picture of
>jd> the Eiffel Tower. But they could just as well have mailed you a
>jd> different second picture that, together with the first, gave a Big
>jd> Mac.

This metaphor may be relevant to the cognitive process by which we DISCOVER an interpretation, but it doesn't apply to the complexity question, which is independent of (or perhaps much bigger than) cognition. If we take the features that make random dots look like the Eiffel Tower versus a Big Mac, those features, and the differences between them, are tiny, compared to the overall number of bits in a random dot array. Besides, to be strictly analogous to the case of the same algorithm formulated in two languages yielding radically different complexities, ALL the random dots would have to be interpretable using either algorithm, whereas the trick with Julesz figures is that only a small subset of the random dots is interpretable (those constituting the figure -- Eiffel Tower or Big Mac, respectively) and not even the same random dots in both cases. (I would also add that the highly constrained class of perceptually ambiguous figures (like the Necker Cube) is more like the rare cases of "dual" interpretability I've already noted.)

>jd> Moreover, it is not true in general that the simpler interpretation is
>jd> always the right one. Someone who is encoding something can arrange
>jd> for there to be a simple interpretation that is incorrect. I suppose
>jd> an example might be where the encrypted form can be decrypted to an
>jd> English text, but the actual message can only be found by taking the
>jd> (English) words that appear after every third word that contains an "a".
>jd>
>jd> Jeff Dalton

Again, this seems to have more to do with the cognitive problem of how to DISCOVER an interpretation than with the question of whether radically different alternative interpretations (for the same symbol system) exist and are accessible in real time. I would also say that the differences in complexity between variant (but coherent) interpretations of the kind you cite here would be tiny and trivial compared to the complexity required to interpret a symbol system after swapping the interpretations of an arbitrary pair of symbol types (such as "if" and "not").

Once you've successfully decrypted something as English, for example, it is trivial to add a second-oder decryption in which a particular message (e.g., in English, or even in French) is embedded after every third word containing an "a." All that would require (if this analogy between algorithms and interpretations is tenable at all) is a few more bits added to the original interpretative algorithm -- which would still leave both algorithms MUCH closer to one another than to the infinite corpus that they both decrypt.

Now there is an analogous argument one might try to make for if/not swapping too: Take the standard English interpretative algorithm and interpret "not" as if it meant "if" and vice versa: Just a few extra bits! But this is not what radical alternative interpretation refers to. It’s not just a matter of using real English, but with the symbols "if" and "not" swapped (i.e., it’s not just a matter of decoding "Not it rains then you can if go out" as "If it rains then you can not go out"). You must have another interpretative algorithm altogether, a "Schmenglish" one, in which the INTERPRETATION of "if" and "not" in standard English strings like "If it rains then you can go out" (plus all the rest of standard English) are given a coherent systematic alternative interpretation in which "if" MEANS "not" and vice versa: A much taller order, and requiring a lot more than a few bits tacked on!

Stevan Harnad

Chaitin, G. (1975) Randomness and mathematical proof. Scientific American 232: 47 - 52.

Rabin, M. O. (1977) Complexity of computations. Communications of the Association of Computer Machinery 20:625-633.

Date: Sun, 12 Apr 1992 07:58:41 -0400 From: Drew McDermott

Let’s distinguish between a computer’s states’ being "microinterpretable" and "macrointerpretable." The former case is what you assume: that if we consider the machine to be a rewrite system, the rewrite rules map one coherently interpretable state into another. Put another way, the rewrite rules specify a change in belief states of the system. By contrast, the states of a macrointerpretable system "sort of line up" with the world in places, but not consistently enough to generate anything like a Tarskian interpretation. What I think you’ve overlooked is that almost all computational processes are at best macrointerpretable.

Take almost any example, a chess program, for instance. Suppose that the machine is evaluating a board position after a hypothetical series of moves. Suppose the evaluation function is a sum of terms. What does each term denote? It is not necessary to be able to say. One might, for instance, notice that a certain term is correlated with center control, and claim that it denotes "the degree of center control," but what does this claim amount to? In many games, the correlation will not hold, and the computer may as a consequence make a bad move. But the evaluation function is "good" if most of the time the machine makes "good moves."

The chess program keeps a tree of board positions. At each node of this tree, it has a list of moves it is considering, and the positions that would result. What does this list denote? The set of moves "worth considering"? Not really; it’s only guessing that these moves are worth considering. We could say that it’s the set the machine "is considering," but this interpretation is trivial.

We can always imose a trivial interpretation on the states of the computer. We can say that every register denotes a number, for instance, and that every time it adds two registers the result denotes the sum. The problem with this idea is that it doesn’t distinguish the interpreted computers from the uninterpreted formal systems, because I can always find such a Platonic universe for the states of any formal system to "refer" to. (Using techniques similar to those used in proving predicate calculus complete.)

More examples: What do the states of a video game refer to? The Mario brothers? Real asteroids?

What do the data structures of an air-traffic control system refer to? Airplanes? What if a blip on the screen is initially the result of thermal noise in the sensors, then tracks a cloud for a while, then switches to tracking a flock of geese? What does it refer to in that case?

Halfway through an application of Newton’s method to an optimization problem involving process control in a factory, what do the various inverted Hessian matrices refer to? Entities in the factory? What in the world would they be? Or just mathematical entities?

If no other argument convinces you, this one should: Nothing prevents a computer from having inconsistent beliefs. We can build an expert system that has two rules that either (a) cannot be interpreted as about medical matters at all; or (b) contradict each other. The system, let us say, happens never to use the two rules on the same case, so that on any occasion its advice reflects a coherent point of view. (Sometimes it sounds like a homeopath, we might say, and sometimes like an allopath.) We would like to say that overall the computer’s inferences and pronouncements are "about" medicine. But there is no way to give a coherent overall medical interpretation to its computational states.

I could go on, but the point is, I hope, clear. For 99.9% of all computer programs, either there is only a trivial intepretation of a program’s state as referring to numbers (or bit strings, or booleans); or there is a vague, unsystematic, error-prone interpretation in terms of the entities the machine is intended to concern itself with. The *only* exceptions are theorem-proving programs, in which these two interpretations coincide. In a theorem prover, intermediate steps are about the same entities as the final result, and the computational rules getting you from step to step are isomorphic to the deductive rules that justify the computational rules. But this is a revealing exception. It’s one of the most pervasive fallacies in computer science to see the formal-systems interpretation of a computer has having some implications for the conclusions it draws when it is interpreted as a reasoning system. I believe you have been sucked in by this fallacy. The truth is that computers, in spite of having trivial interpretations as deductive systems, can be used to mimic completely nondeductive systems, and that any semantic framework they approximate when viewed this way will bear no relation to the low-level deductive semantics.

I suspect Searle would welcome this view, up to a point. It lends weight to his claim that semantics are in the eye of the beholder. One way to argue that an air-traffic control computer’s states denote airplanes is to point out that human users find it useful to interpret them this way on almost every occasion. However, the point at issue right now is whether semantic interpretability is part of the definition of "computer." I argue that it is not; a computer is what it is regardless of how it is interpreted. I buttress that observation by pointing out just how unsystematic most interpretations of a computer’s states are. However, if I can win the argument about whether computers are objectively given, and uninterpreted, then I can go on to argue that unsystematic interpretations of their states can be objectively given as well.

-- Drew McDermott

From: Stevan Harnad

Drew McDermott wrote:

>dm> Let's distinguish between a computer's states' being
>dm> "microinterpretable" and "macrointerpretable." The former case is what
>dm> you assume: that if we consider the machine to be a rewrite system, the
>dm> rewrite rules map one coherently interpretable state into another. Put
>dm> another way, the rewrite rules specify a change in belief states of the
>dm> system. By contrast, the states of a macrointerpretable system "sort of
>dm> line up" with the world in places, but not consistently enough to
>dm> generate anything like a Tarskian interpretation. What I think you've
>dm> overlooked is that almost all computational processes are at best
>dm> macrointerpretable.

Drew, you won't be surprised by my immediate objection to the word "belief" above: Until further notice, a computer has physical states, not belief states, although some of those physical states might be interpretable -- whether "macro" or "micro" I'll get to in a moment -- AS IF they were beliefs. Let's pretend that's just a semantic quibble (it's not, of course, but rather a symptom of hermeneutics creeping in; however, let's pretend).

You raise four semi-independent issues:

(1) Does EVERY computer implementing a program have SOME states that are interpretable as referring to objects, events and states of affairs, the way natural language sentences are?

(2) Are ALL states in EVERY computer implementing a program interpretable as referring... (etc.)?

(3) What is the relation of such language-like referential interpretability and OTHER forms of interpretability of states of a computer implementing a program?

(4) What is the relation of (1) - (3) to the software hierarchy, from hardware, to machine-level language, to higher-level compiled languages, to their English interpretations?

My answer would be that not all states of a computer implementing a program need be interpretable, and not all the interpretable states need be language-like and about things in the world (they could be interpretable as performing calculations on numbers, etc.), but ENOUGH of the states need to be interpretable SOMEHOW, otherwise the computer is just performing gibberish (and that's usually not what we use computers to do, nor do we describe them as such), and THAT's the interpretability that's at issue here.

Some of the states may have external referents, some internal referents (having to do with the results of calculations, etc.). And there may be levels of interpretation, where the higher-level compiled languages have named "chunks" that are (macro?)interpretable as being about objects, whereas the lower-level languages are (micro?)interpretable only as performing iterative operations, comparisons, etc. Although it's easy to get hermeneutically lost in it, I think the software hierarchy, all the way up to the highest "virtual machine" level, does not present any fundamental mysteries at all. Low-level operations are simply re-chunked at a higher level so more general and abstract computations can be performed. I can safely interpret a FORTRAN statement as multiplying 2 x 2 without worrying about how that's actually being implemented at the machine-language or hardware level -- but it IS being implemented, no matter how complicated the full hardware story for that one operation would be.

>dm> Take almost any example, a chess program, for instance. Suppose that
>dm> the machine is evaluating a board position after a hypothetical series
>dm> of moves. Suppose the evaluation function is a sum of terms. What does
>dm> each term denote? It is not necessary to be able to say. One might, for
>dm> instance, notice that a certain term is correlated with center control,
>dm> and claim that it denotes "the degree of center control," but what does
>dm> this claim amount to? In many games, the correlation will not hold, and
>dm> the computer may as a consequence make a bad move. But the evaluation
>dm> function is "good" if most of the time the machine makes "good moves."

I'm not sure what an evaluation function is, but again, I am not saying every state must be interpretable. Even in natural language there are content words (like "king" and "bishop") that have referential interpretations and function words ("to" and "and") that have at best only syntactic or functional interpretations. But some of the internal states of a chess-plying program surely have to be interpretable as referring to or at least pertaining to chess-pieces and chess-moves, and those are the ones at issue here. (Of course, so are the mere "function" states, because they too will typically have something to do with (if not chess then) calculation, and that's not gibberish either.

>dm> The chess program keeps a tree of board positions. At each node of this
>dm> tree, it has a list of moves it is considering, and the positions that
>dm> would result. What does this list denote? The set of moves "worth
>dm> considering"? Not really; it's only guessing that these moves are worth
>dm> considering. We could say that it's the set the machine "is
>dm> considering," but this interpretation is trivial.

And although I might make that interpretation for convenience in describing or debugging the program (just as I might make the celebrated interpretation that first got Dan Dennett into his "intentional stance," namely, that "the computer thinks it should get it's queen out early"), I would never dream of taking such interpretations literally: Such high level mentalistic interpretations are simply the top of the as-if hierarchy, a hierarchy in which intrinsically meaningless squiggles and squoggles can be so interpreted that (1) they are able to bear the systematic weight of the interpretation (as if they "meant" this, "considered/believed/thought" that, etc.), and (2) the interpretations can be used in (and even sometimes hard-wired to) the real world (as in interpreting the squiggles and squoggles as pertaining to chess-men and chess-moves).

>dm> We can always impose a trivial interpretation on the states of the
>dm> computer. We can say that every register denotes a number, for
>dm> instance, and that every time it adds two registers the result denotes
>dm> the sum. The problem with this idea is that it doesn't distinguish the
>dm> interpreted computers from the uninterpreted formal systems, because I
>dm> can always find such a Platonic universe for the states of any formal
>dm> system to "refer" to. (Using techniques similar to those used in

>dm> proving predicate calculus complete.)

I’m not sure what you mean, but I would say that whether they are scratches on a paper or dynamic states in a machine, formal symbol systems are just meaningless squiggles and squoggles unless you project an interpretation (e.g., numbers and addition) onto them. The fact that they will bear the systematic weight of that projection is remarkable and useful (it’s why we’re interested in formal symbol systems at all), but certainly not evidence that the interpretation is intrinsic to the symbol system; it is only evidence of the fact that the system is indeed a nontrivial symbol system (in virtue of the fact that it is systematically interpretable). Nor (as is being discussed in other iterations of this discussion) are coherent, systematic "nonstandard" alternative interpretations of formal symbol systems that easy to come by.

>dm> More examples: What do the states of a video game refer to? The Mario
>dm> brothers? Real asteroids?

They are interpretable as pertaining (not referring, because there’s no need for them to be linguistic) to (indeed, they are hard-wireable to) the players and moves in the Mario Brothers game, just as in chess. And the graphics control component is interpretable as pertaining to (and hard-wireable to the bit-mapped images of) the icons figuring in the game. A far cry from uninterpretable squiggles and squoggles.

>dm> What do the data structures of an air-traffic control system refer to?
>dm> Airplanes? What if a blip on the screen is initially the result of
>dm> thermal noise in the sensors, then tracks a cloud for a while, then
>dm> switches to tracking a flock of geese? What does it refer to in that
>dm> case?

I don’t know the details, but I’m sure a similar story can be told here: Certain squiggles and squoggles are systematically interpretable as signaling (and mis-signaling) the presence of an airplane, and the intermediate calculations that lead to that signaling are likewise interpretable in some way. Running computer programs are, after all, not black boxes inexplicably processing input and output. We design them to do certain computations; we know what those computations are; and what makes them computations rather than gibberish is that they are interpretable.

>dm> Halfway through an application of Newton’s method to an optimization
>dm> problem involving process control in a factory, what do the various
>dm> inverted Hessian matrices refer to? Entities in the factory? What in
>dm> the world would they be? Or just mathematical entities?

The fact that the decomposition is not simple does not mean that the intermediate states are all or even mostly uninterpretable.

>dm> If no other argument convinces you, this one should: Nothing prevents
>dm> a computer from having inconsistent beliefs. We can build an expert
>dm> system that has two rules that either (a) cannot be interpreted as
>dm> about medical matters at all; or (b) contradict each other. The system,
>dm> let us say, happens never to use the two rules on the same case, so
>dm> that on any occasion its advice reflects a coherent point of view.
>dm> (Sometimes it sounds like a homeopath, we might say, and sometimes like

>dm> an allopath.) We would like to say that overall the computer's
>dm> inferences and pronouncements are "about" medicine. But there is no way
>dm> to give a coherent overall medical interpretation to its computational
>dm> states.

I can't follow this: The fact that a formal system is inconsistent, or can potentially generate inconsistent performance, does not mean it is not coherently interpretable: it is interpretable as being inconsistent, but as yielding mostly correct performance nevertheless. [In other words, "coherently interpretable" does not mean "interpretable as coherent" (if "coherent" presupposes "consistent").]

And, ceterum sentio, the system has no beliefs; it is merely systematically interpretable as if it had beliefs (and inconsistent ones, in this case). Besides, since even real people (who are likewise systematically interpretable, but not ONLY systematically interpretable: also GROUNDED by their TTT-powers in the real world) can have inconsistent real beliefs, I'm not at all sure what was meant to follow from your example.

>dm> I could go on, but the point is, I hope, clear. For 99.9% of all
>dm> computer programs, either there is only a trivial interpretation of a
>dm> program's state as referring to numbers (or bit strings, or booleans);
>dm> or there is a vague, unsystematic, error-prone interpretation in terms
>dm> of the entities the machine is intended to concern itself with. The
>dm> *only* exceptions are theorem-proving programs, in which these two
>dm> interpretations coincide. In a theorem prover, intermediate steps are
>dm> about the same entities as the final result, and the computational
>dm> rules getting you from step to step are isomorphic to the deductive
>dm> rules that justify the computational rules. But this is a revealing
>dm> exception. It's one of the most pervasive fallacies in computer science
>dm> to see the formal-systems interpretation of a computer as having some
>dm> implications for the conclusions it draws when it is interpreted as a
>dm> reasoning system. I believe you have been sucked in by this fallacy.
>dm> The truth is that computers, in spite of having trivial interpretations
>dm> as deductive systems, can be used to mimic completely nondeductive
>dm> systems, and that any semantic framework they approximate when viewed
>dm> this way will bear no relation to the low-level deductive semantics.

My view puts no special emphasis on logical deduction, nor on being interpretable as doing logical deduction. Nor does it require that a system be interpretable as if it had only consistent beliefs (or any beliefs at all, for that matter). It need be interpretable only in the way symbol strings in English, arithmetic, C or binary are interpretable.

>dm> I suspect Searle would welcome this view, up to a point. It lends
>dm> weight to his claim that semantics are in the eye of the beholder.
>dm> One way to argue that an air-traffic control computer's states denote
>dm> airplanes is to point out that human users find it useful to
>dm> interpret them this way on almost every occasion. However, the point
>dm> at issue right now is whether semantic interpretability is part of the
>dm> definition of "computer." I argue that it is not; a computer is what
>dm> it is regardless of how it is interpreted. I buttress that

>dm> observation by pointing out just how unsystematic most interpretations
>dm> of a computer's states are. However, if I can win the argument about
>dm> whether computers are objectively given, and uninterpreted, then I
>dm> can go on to argue that unsystematic interpretations of their states
>dm> can be objectively given as well.
>dm>
>dm> -- Drew McDermott

If you agree with Searle that computers can't be distinguished from non-computers on the basis of interpretability, then I have to ask you what (if anything) you DO think distinguishes computers from non-computers? Because "Everything is a computer" would simply eliminate (by fiat) the substance in any answer at all to the question "Can computers think?" (or any other question about what can or cannot be done by a computer, or computationally). Some in this discussion have committed themselves to universality and a complexity-based criterion (arbitrary rival interpretations are NP-complete). Where do you stand?

Stevan Harnad

From: Brian C Smith Date: Thu, 16 Apr 1992 11:43:32 PDT

I can't help throwing in a number of comments into this discussion:

1) ON UNIVERSALITY: All metrics of equivalence abstract away from certain details, and focus on others. The metrics standardly used to show universality are extraordinarily coarse-grained. They are (a) essentially behaviourist, (b) blind to such things as timing, and (c) (this one may ultimately matter the most), promiscuous exploiters of implementation, modelling, simulation, etc. Not only does it strike me as extremely unlikely that (millenial versions of) "cognitive", "semantic", etc., will be this coarse-grained, but the difference between a model and the real thing (ignored in the standard equivalence metrics) is exactly what Searle and others are on about. It therefore does not follow, if X is cognitive, and Y provably equivalent to it (in the standard theoretic sense), that Y is cognitive.

This considerations suggest not only that universality may be of no particular relevance to cognitive science, but more seriously that it is somewhere between a red herring and a mine field, and should be debarred from arguments of cognitive relevance.

2) ON ORIGINAL INTENTIONALITY: Just a quick one. In some of the notes, it seemed that *intrinsic* and *attributed* were being treated as opposites. This is surely false. Intrinsic is presumably opposed to something like extrinsic or relational. Attributed or observer- supplied is one particular species of relational, but there are many others. Thus think about the property of being of average height. This property doesn't inhere within an object, but that doesn't make it ontologically dependent on observation or attribution (at least no more so that anything else [cf. Dietrich]).

There are lots of reasons to believe that semantics, even original semantics, will be relational. More seriously, it may even be that our *capacity* for semantics is relational (historical, cultural, etc. -- this is one way to understand some of the deepest arguments that language is an inexorably cultural phenomenon). I.e., it seems to me a mistake to assume that *our* semantics is intrinsic in us. So arguing that computers' semantics is not intrinsic doesn't cut it as a way to argue against computational cognitivism.

3) ON FORMAL SYMBOL MANIPULATION: In a long analysis (20 years late, but due out soon) I argue that actual, real-world computers are not formal symbol manipulators (or, more accurately, that there is no coherent reading of the term "formal" under which they are formal). Of many problems, one that is relevant here is that the inside/ outside boundary does not align with the symbol/referent boundary -- a conclusion that wreaks havoc on traditional notions of transducers, claims of the independence of syntax and semantics, the relevance of "brain in a vat" thought experiments, etc.

4) ON THE "ROBOTIC" SOLUTION: Imagine someone trying to explain piano music by starting with the notion of a melody, then observing that more than one note is played at once, and then going on to say that there must also be chords. Maybe some piano music can be described like that: as melody + chords. But not a Beethoven sonata. The consequence of "many notes at once" is not that one *adds* something (chords) to the prior idea of a single-line melody. Once you've got the ability to have simultaneous notes, the whole ball game changes.

I worry that the robotic reply to Searle suffers the same problem. There's something right about the intuition behind it, having to do with real-world engagement. But when you add it, it is not clear whether the original notion (of formal symbol manipulation, or even symbol manipulation at all) survives, let alone whether it will be a coherent part of the expanded system. I.e., "symbol + robotic grounding" seems to me all too similar to "melody + chords".

If this is true, then there is a very serious challenge as to what notions *are* going to explain the expanded "engaged with the real world" vision. One question, the one on the table, is whether or not they will be computational (my own view is: *yes*, in the sense that they are exactly the ones that are empirically needed to explain Silicon Valley practice; but *no*, in that they will neither be an extension to nor modification of the traditional formal symbol manipulation construal, but will instead have to be redeveloped from scratch). More serious than whether they are computational, however, is what those notions *will actually be*. I don't believe we know.

5) ON TYPES: On March 22, Gary Hatfield raised a point whose importance, I believe, has not been given its due. Over the years, there have been many divisions and distinctions in AI and cognitive science: neat vs. fuzzy; logicist vs. robotic; situated vs. non-situated; etc. I have come to believe, however, that far and away the most important is whether people assume that the TYPE STRUCTURE of the world can be taken as explanatorily and unproblematically given, or whether it is something that a theory of cognition/computation /intentionality/etc. must explain. If you believe that the physical characterisation of a system is given (as many writers seem to do), or that the token characterisation is given (as Haugeland would lead us to believe), or that the set of states is given (as Chalmers seems to), or that the world is parsed in advance (as set theory & situation theory both assume), then many of the foundational questions don't seem to be all that problematic.

Some of us, however, worry a whole lot about where these type structures come from. There is good reason to worry: it is obvious, once you look at it, that the answers to all the interesting questions come out different, if you assume different typing. So consider the disussions of physical implementation. Whether there is a mapping of physical states onto FSA states depends on what you take the physical and FSA states to be. Not only that, sometimes there seems to be no good reason to choose between different typings. I once tried to develop a theory of representation, for example, but it had the unfortunate property that the question of whether maps were isomorphic representations of territory depended on whether I took the points on the maps to be objects, and the lines to be relations between them, or took the lines to be objects and the points to be relations (i.e., intersections) between *them*. I abandoned the whole project, because it was clear that something very profound was wrong: my analysis depended far too much on my own, inevitably somewhat arbitrary, theoretic decisions. I, the theorist, was implicitly, and more or less unwittingly, *imposing* the structure of the solution to my problem onto the subject matter beforehand.

Since then, I have come to believe that explaining the rise of ontology (objects, properties, relations, types, etc.) is part and parcel of giving an adequate theory of cognition. It's tough sledding, and this is not the place to go into it. But it is important to get the issue of whether one believes that one can assume the types in advance out onto the table, because I think implicit disagreement over this almost methodological issue can subvert communication on the main problems of the day.

Brian Smith

(P.S.: Is there a reason not to have a mailing list that each of us can post to directly?)

[The symbol grounding list is not an unmoderated list; it is moderated by me. I post all substantive messages, but if it were unmoderated it would quickly degenerate into what goes on on comp.ai. -- SH]

Date: Sat, 18 Apr 92 17:29:50 MDT To: mcdermott-drew@CS.YALE.EDU Cc: harnad%Princeton.EDU.hayes@cs.stanford.edu

Drew, clearly you have an antisemantic axe to grind, but its not very sharp.

First of all, of course you are right that many computational processes don't have a constant coherent interpretation. But not 99% of them. Let's look at your examples. First the chess program's list of moves. That this list denotes any list of chess moves - that is, moves of actual chess - is already enough of an interpretation to be firmly in the world of intentionality. You might ask, what IS a move of actual chess, and I wouldn't want to have to wait for a philosopher's answer, but the point here is that it certainly isn't something inside a computer: some kind of story has to be told in which that list denotes (or somehow corresponds to, or has as its meaning) something other than bit-strings. And this kind of story is an essential part of an account of, for example, the correctness of the chess-playing code. Your point that the heuristics which choose a particular set of moves (or which assign particular values of some evaluation function to a move) are in some sense ill-defined is correct, but that is not to say they are uninterpretable. A bitstring has many interpretations which are not numbers and have nothing at all to do with chess, so to claim that these are the meanings is to say something significant.

Suppose I were to build a machine which treated its bit-strings like base-1 integers, so that N was represented by a consecutive string of ones N long. Now your interpretation of addition will fail. So it isn't completely trivial.

Consider again the air traffic control system which gets confused by thermal noise, then clouds, then geese. This is a familiar situation, in which a knower has confused knowledge. But the model theory accounts for this perfectly well. Its beliefs were false, poor thing, but they had content: it thought there was an airplane there. To give a proper account of this requires the use of modality and a suitable semantics for it, as I know you know. One has to say something like its blip denoted an airplane in the possible worlds consistent with its beliefs. But look, all this is semantics, outside the formal syntactic patterns of its computational memory. And just CALLING it an "air-traffic control system" implies that its computational states have some external content.

Your inconsistent-beliefs point misses an important issue. If that expert system has some way of ensuring that these contradictory rules never meet, then it has a consistent interpretation, trivially: we can regard the mechanism which keeps them apart as being an encoding of a syntactic difference in its rule-base which restores consistency. Maybe one set of rules is essentially written with predicates with an "allo-" prefix and the others with a "homeo-". You might protest that this is cheating, but I would claim not: in fact, we need a catalog of such techniques for mending consistency in sets of beliefs, since people seem to have them and use them to 'repair' their beliefs constantly, and making distinctions like this is one of them (as in, "Oh, I see, must be a different kind of doctor"). If on the other hand the system has no internal representation of the distinction, even implici t, but just happens to never bring the contradiction together, then it is in deep trouble as it will soon just happen to get its knowledge base into total confusion. But in any case, it is still possible to interpret an inconsistent set of beliefs as meaningful, since subsets of it are. We might say of this program, as we sometimes do of humans, that it was confused, or it seemed to keep changing its mind about treatment procedures: but this is still ABOUT medicine. A very naive application of Tarskian models to this situation would not capture the necessary subtlety of meaning, but that doesn't make it impossible.

Finally, there is no need to retreat to this idea of the interpretation being a matter of human popularity. The reason the states of an autopilot denote positions of the airplane is not because people find it useful to interpret them that way, but because (with very high probability) the airplane goes where it was told to.

Pat Hayes

Date: Mon, 20 Apr 92 09:58:09 EDT From: "Stevan Harnad"

bd> Date: Sun, 19 Apr 92 20:06:37 PDT bd> From: dambrosi@research.CS.ORST.EDU (Bruce Dambrosio) bd> bd> Stevan: bd> bd> I am puzzled by one thing, which perhaps was discussed earlier: why do bd> you, of all people, believe that a definition satisfying your bd> requirements might exist? This seems to me quite a quixotic quest. bd> bd> A definition is a symbolically specified mapping from the objects bd> denoted by some set of symbols to the objects denoted by the symbol bd> being defined. But if, as you claim, the process by which relationship bd> is established (grounding) is such that it cannot be adequately bd> described symbolically (I take this to be the heart of the symbol bd> grounding position), then how can one ever hope to describe the bd> relationship between two groundings symbolically? At best, one can only bd> hope for a rough approximation that serves to guide the hearer in the bd> right direction. I may know what a computer is, but be quite unable to bd> give you a definition that stands up to close scrutiny. Indeed, such a bd> situation would seem to be evidence in favor of symbol grounding as a bd>

significant issue. Am I naive or has this already been discussed? bd> bd> Bruce D’Ambrosio

Bruce,

This has not been explicitly discussed, but unfortunately your description of the symbol grounding problem is not quite correct. The problem is not that we cannot give adequate definitions symbolically (e.g., linguistically); of course we can: We can define adequately anything, concrete or abstract, that we understand adequately enough to define.

The symbol grounding problem is only a problem for those (like mind modelers) who are trying to design systems in which the meanings of the symbols are INTRINSIC to the system, rather than having to be mediated by our (grounded) interepretations. There is nothing whatsoever wrong with ungrounded symbol systems if we want to use them for other purposes, purposes in which our interpretations are free to mediate. A dictionary definition is such a mediated use, and it does not suffer from the symbol grounding problem. The example of the Chinese-Chinese Dictionary-Go-Round that I described in Harnad (1990) was one in which the dictionary was being used by someone who knew no Chinese! For him the ungroundedness of the dictionary (and the fact that its use cannot be mediated by his own [nonexistent] grounded understanding of Chinese) is indeed a problem, but not for a Chinese speaker.

If our notions of "computer" and "computation" are coherent ones (and I suspect they are, even if still somewhat inchoate) then there should be no more problem with defining what a computer is than in defining what any other kind of object, natural or artificial, is. The alternatives (that everything is a computer, or everything is a computer to some degree, or nothing is a computer), if they are the correct ones, would mean that a lot of the statements we make in which the word "computer" figures (as in "computers can/cannot do this/that") would be empty, trivial, or incoherent.

One pass at defining computation and computer would be as, respectively, syntactic symbol manipulation and universal syntactic symbol manipulator, where a symbol system is a set of objects (symbols) that is manipulated according to (syntactic) rules operating only on their shapes (not their meanings), the symbols and symbol manipulations are systematically interpretable as meaning something (and the interpretation is cryptologically nontrivial), but the shapes of the elementary symbol tokens are arbitrary in relation to what they can be interpreted as meaning. I, for one, could not even formulate the symbol grounding problem if there were no way to say what a symbol system was, or if everything was a symbol system.

As to the question of approximate grounding: I discuss this at length in Harnad (1987). Sensory groundings are always provisional and approximate (because they are relative to the sample of confusable alternatives encountered to date). Definitions may be provisional and empirical ones, or they may be stipulative and analytical. If the latter, they are not approximate, but exact "by definition." I would argue, however, that even high-level exact definitions depend for our understanding on the grounding of their symbols in lower-level symbols, which are in turn grounded ultimately in sensory symbols (which are indeed provisional and approximate). This just suggests that symbol grounding should not be confused with ontology.

There are prominent philosophical objections to this kind of radical bottom-uppism, objections of which I am quite aware and have taken some passes at answering (Harnad 1992). The short answer is that bottom-uppism cannot be assessed by introspective analysis alone and has never yet been tried empirically; in particular, no one knows HOW we actually manage to sort, label and describe objects, events and states of affairs as we do, but we can clearly do it; hence, until further notice, input information (whether during our lifetimes or during the evolutionary past that shaped us) is the only candidate source for this remarkable capacity.

Stevan Harnad

Harnad, S. (1987) The induction and representation of categories. In: In: S. Harnad (ed.) Categorical Perception: The Groundwork of Cognition. New York: Cambridge University Press.

Harnad, S. (1990) The Symbol Grounding Problem. Physica D 42: 335-346.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Date: Tue, 21 Apr 92 18:22:32 EDT From: "Stevan Harnad"

ARE "GROUNDED SYMBOL SYSTEMS" STILL SYMBOL SYSTEMS?

> Brian C Smith wrote:

>bs> It... does not follow [that] if X is cognitive, and Y provably
>bs> equivalent to it (in the standard theoretic sense), that Y is
>bs> cognitive.

Of course not; in fact one wonders why this even needs to be said! Equivalence "in the standard sense" is computational equivalence, not physical-causal equivalence. Whether someone is drowned in water or in beer is equivalent insofar as drowning is concerned, because the drowning is real in both cases. But if the drowning is "virtual" (i.e., a computer-simulated person is "drowned" in computer-simulated water) there is no drowning at all going on, no matter how formally equivalent the symbols may be to real drowning.

>bs> In some of the notes, it seemed that *intrinsic* and *attributed* were
>bs> being treated as opposites. This is surely false. Intrinsic is
>bs> presumably opposed to something like extrinsic or relational.
>bs> Attributed or observer-supplied is one particular species of
>bs> relational, but there are many others.

I’ve never understood why so much emphasis is placed by philosophers on the difference between monadic ("intrinsic") and polyadic ("relational") properties. Surely that’s not the real issue in mind modeling. What we want is that symbols should mean X not just because we interpret them as meaning X but because they (also) mean X independently of our interpretations. Their meaning has to be autonomously GROUNDED in something other than just their being able to bear the systematic weight of our interpretations.

The string of symbols "the cat is on the mat," whether it is instantiated on the inert pages of a book or as a dynamic state in a computer running a LISP program, is systematically interpretable as meaning "the cat is on the mat" (in relation to the rest of the symbol system) but it does not mean "the cat is on the mat" on its own, autonomously, the way I do when I think and mean "the cat is on the mat," because I, unlike the book or the computer, don't mean "the cat is on the mat" merely in virtue of the fact that someone else can systematically interpret me as meaning that.

So the real problem is how to ground meaning autonomously, so as not to leave it hanging from a skyhook of mere interpretation or interpretability. The solution may still turn out to be "relational," but so what? According to my own robotic grounding proposal, for example, a robot's symbols would have autonomous meaning (or, to be noncommittal, let's just say they would have autonomous "grounding") because their use would be governed and constrained by whatever it takes to make the robot capable of interacting TTT-indistinguishably with the very objects to which its symbols were interpretable as referring. The meaning of the robot's symbols is grounded in its robotic capacity instead of depending only on how the symbols can be or actually are interpreted by us. But note that this is merely a case of one set of "relations" (symbol/symbol relations and their interpretations) being causally constrained to be coherent with another set of "relations" (symbol/object relations in the world).

The source, I think, of the undue preoccupation with monadic properties is the (correct) intuition that our thoughts are meaningful in and of themselves, not because of how their interrelations are or can be interpreted by others. Probably the fact that all thoughts are the thoughts of a conscious subject (and that their meaning is a meaning to that conscious subject) also contributed to the emphasis on the autonomy and "intrinsic" nature of meaning.

>bs> There are lots of reasons to believe that semantics, even original
>bs> semantics, will be relational... it may even be that our
>bs> *capacity* for semantics is relational (historical, cultural, etc)...
>bs> it seems to me a mistake to assume that *our* semantics is intrinsic in
>bs> us. So arguing that computers' semantics is not intrinsic doesn't cut
>bs> it as a way to argue against computational cognitivism.

To agree that the meanings of the symbols inside a robot are grounded in (say) the robot's actual relations to the objects to which its symbols can be interpreted as referring is still not to agree that the locus of those meanings is any wider -- in either time or space -- than the robot's body (which includes the projections and effects of real world objects on its sensorimotor surfaces).

>bs> [In a forthcoming paper] I argue that actual, real-world computers are
>bs> not formal symbol manipulators (or, more accurately, that there is no
>bs> coherent reading of the term "formal" under which they are formal).
>bs> Of many problems, one that is relevant here is that the inside/
>bs> outside boundary does not align with the symbol/referent boundary --
>bs> a conclusion that wreaks havoc on traditional notions of transducers,
>bs> claims of the independence of syntax and semantics, the relevance of
>bs> "brain in a vat" thought experiments, etc.

One would have to see this forthcoming paper, but my intuition is that a lot of red herrings have been and continue to be raised whenever one attempts to align (1) the internal/external distinction for a physical system with (2) what is going on "inside" or "outside" a mind. The first. I think, is largely unproblematic: We can safely (though not always usefully) distinguish the inside and the outside of a computer or a robot, as well as the I/O vs. the processing of a symbol system. What is inside and outside a mind is another story, one that I think is incommensurable with anything but the grossest details of the physical inside/outside story.

As a first pass at "formal," how about: A symbol system consists of a set of objects (elementary symbols and composite symbols) plus rules for manipulating the symbols. The rules operate only on the physical shapes of the symbols, not their meanings (and the shapes of the elementary symbols are arbitrary), yet the symbols are systematically interpretable as meaning something. The rules for manipulating the symbols on the basis of their shapes are called "syntactic" or "formal" rules.

A computer is a dynamical system that mechanically implements the symbols, the symbol manipulations, and the rules (constraints) governing the symbol manipulations.

>bs> Imagine someone trying to explain piano music by starting with the
>bs> notion of a melody, then observing that more than one note is played at
>bs> once, and then going on to say that there must also be chords. Maybe
>bs> some piano music can be described like that: as melody + chords. But
>bs> not a Beethoven sonata. The consequence of "many notes at once" is not
>bs> that one *adds* something (chords) to the prior idea of a single-line
>bs> melody. Once you've got the ability to have simultaneous notes, the
>bs> whole ball game changes.

Some symbols can be indirectly grounded this way, using propositions with symbols that either have direct sensory grounding or are near to their sensory grounding (e.g., "A ‘zebra’ is a horse with stripes"), but many symbols cannot be adequately grounded by symbolic description alone and require direct sensory acquaintance. This is just more evidence for the importance of sensorimotor grounding.

>bs> I worry that the robotic reply to Searle suffers the same problem.
>bs> There's something right about the intuition behind it, having to do
>bs> with real-world engagement. But when you add it, it is not clear
>bs> whether the original notion (of formal symbol manipulation, or even
>bs> symbol manipulation at all) survives, let alone whether it will be a
>bs> coherent part of the expanded system. I.e., "symbol + robotic
>bs> grounding" seems to me all too similar to "melody + chords".

The standard robot reply to Searle is ineffectual, because it retains the (symbols-only) Turing Test (TT) as the crucial test for having a mind and simply adds on arbitrary peripheral modules to perform robotic functions. My own "robot" reply (which I actually call the "Total" reply) rejects the TT altogether for the "Total Turing Test" (TTT) and is immune to Searle's argument because the TTT cannot be passed by symbol manipulation alone, and Searle (on pain of the "System Reply," which normally fails miserably, but not in the case of the TTT) can fully implement only pure implementation-independent symbol manipulation, not implementation-dependent nonsymbolic processes such as transduction, which are essential for passing the TTT.

On the other hand, I agree that grounded "symbol systems" may turn out to be so radically different from pure symbol systems as to make it a different ballgame altogether (the following passage is from the Section entitled "Analog Constraints on Symbols" in Harnad 1992):

"Recall that the shapes of the symbols in a pure symbol system are arbitrary in relation to what they stand for. The syntactic rules, operating on these arbitrary shapes, are the only constraint on the manipulation of the symbols. In the kind of hybrid system under consideration here, however, there is an additional source of constraint on the symbols and their allowable combinations, and that is the nonarbitrary shape of the categorical representations that are "connected" to the elementary symbols: the sensory invariants that can pick out the object to which the symbol refers on the basis of its sensory projection. The constraint is bidirectional. The analog space of resemblances between objects is warped in the service of categorization -- similarities are enhanced and diminished in order to produce compact, reliable, separable categories. Objects are no longer free to look quite the same after they have been successfully sorted and labeled in a particular way. But symbols are not free to be combined purely on the basis of syntactic rules either. A symbol string must square not only with its syntax, but also with its meaning, i.e., what it, or the elements of which it is composed, are referring to. And what they are referring to is fixed by what they are grounded in, i.e., by the nonarbitrary shapes of the iconic projections of objects, and especially the invariants picked out by the neural net that has accomplished the categorization."

>bs> If this is true, then there is a very serious challenge as to what
>bs> notions *are* going to explain the expanded "engaged with the real
>bs> world" vision. One question, the one on the table, is whether or not
>bs> they will be computational (my own view is: *yes*, in the sense that
>bs> they are exactly the ones that are empirically needed to explain
>bs> Silicon Valley practice; but *no*, in that they will neither be an
>bs> extension to nor modification of the traditional formal symbol
>bs> manipulation construal, but will instead have to be redeveloped from
>bs> scratch). More serious than whether they are computational, however, is
>bs> what those notions *will actually be*. I don't believe we know.

I think I agree: The actual role of formal symbol manipulation in certain dedicated symbol systems (e.g., TTT-scale robots) may turn out to be so circumscribed and/or constrained that the story of the constraints (the grounding) will turn out to be more informative than the symbolic story.

>bs> the most important [distinction in AI and cognitive science] is whether
>bs> people assume that the TYPE STRUCTURE of the world can be taken as
>bs> explanatorily and unproblematically given, or whether it is something
>bs> that a theory of cognition/computation /intentionality/etc. must
>bs> explain. If you believe that the physical characterisation of a system
>bs> is given (as many writers seem to do), or that the token
>bs> characterisation is given (as Haugeland would lead us to believe), or
>bs> that the set of states is given (as Chalmers seems to), or that the
>bs> world is parsed in advance (as set theory & situation theory both
>bs> assume), then many of the foundational questions don't seem to be all
>bs> that problematic... Some of us, however, worry a whole lot about where
>bs> these type structures come from... [E]xplaining the rise of ontology
>bs> (objects, properties, relations, types, etc.) is part and parcel of
>bs> giving an adequate theory of cognition.
>bs>
>bs> Brian Smith

It is for this reason that I have come to believe that categorical perception and the mechanisms underlying our categorizaing capacity are the groundwork of cognition.

Stevan Harnad

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Date: Tue, 21 Apr 92 20:34:02 EDT From: "Stevan Harnad"

From: Pat Hayes Date: Wed, 15 Apr 92 16:02:09 MDT

>sh> ...if a mind supervenes on (the right)
>sh> computations because of their computational
>sh> properties (rather than because of the physical
>sh> details of any particular implementation of
>sh> them), then it must supervene on ALL
>sh> implementations of those computations. I think
>sh> Searle's Chinese Room Argument has successfully
>sh> pointed out that this will not be so ...

>ph> No, only if you believe that what Searle in the room is doing is
>ph> letting a program run on him, which I think is clearly false. Searle's
>ph> Chinese Room argument doesn't SHOW anything. It can be used to bolster
>ph> a belief one might have about computations, but if one doesn't accept
>ph> that as a premise, than it doesn't follow as a conclusion either. The
>ph> "argument" is just an intuition pump, as Dennett observed a decade
>ph> ago.

Pat, "intuition pump" is not a pejorative, if it pumps true. I will be happy to consider the implications of the fact that Searle, doing everything the computer does, does not count as a valid implementation of the same computer program -- as soon as you specify and argue for what you mean by implementation and why Searle's would not qualify. Until then, I don't see why EVERY system that processes the same symbols, follows the same (syntactic) rules and steps through the same states doesn't qualify as a valid implementation of the same program.

>sh> transducers, for example, are no more
>sh> implementation-independent than digestion is.

>ph> Well, I see what you mean and agree, but one does have to be careful.
>ph> The boundary of implementation-independence can be taken very close to
>ph> the skin. For example, consider a robot with vision and imagine
>ph> replacing its tv cameras with more modern ones which use an array of
>ph> light-sensitive chips rather than scanning something with an electron
>ph> beam. It really doesn't matter HOW it works, how the physics is
>ph> realised, provided it sends the right signals back along its wires. And
>ph> this functional specification can be given in terms of the physical
>ph> energies which are input to it and the syntax of its output. So we are

>ph> in supervenience from the skin inwards.

I agree that the layer between the "shadow" that objects cast on our transducers and the symbols they are cashed into at the very next layer could in principle be VERY thin -- if indeed the rest of the story were true, which is that the signals are just hurtling headlong toward a symbolic representation. However, I don't believe the rest of the story is true! I think most of the brain is preserving sensory signals in various degrees of analog form (so we would probably do well to learn from this). In fact, I think it's as likely that a mind is mostly symbolic, with just a thin analog layer mediating input and output to the world, as that a plane or a furnace are mostly symbolic, with a thin analog layer mediating input and output.

But even if the transducer layer WERE that thin, my point would stand (and that thin layer would then simply turn out to be critically important for the implementation of mental states). Although I don't get much insight from the concept of "supervenience," it would be the analog-plus-symbolic system on which mental states would "supervene," not the symbolic part alone, even if the analog layer was only one micron thick.

I AM a functionalist about analog systems though. There's more than one way to skin an analog cat: As long as devices are analog, and support the same I/O, they don't have to be physically identical: The omatidia of the horseshoe crab transduce light just as literally as mammalian retinae (or synthetic optical transducers) do; as long as they really transduce light and generate the same I/O, they're functionally equivalent enough for me, as transducers. Same is true for internal A/A transforms, with retinal signals that code light in the intensity domain going into some other continuous variable (or even A/D into the frequency domain) as long as they are functionally equivalent and invertible at the I/O end.

>ph> This is like my point about language. While I think you are ultimately
>ph> correct about the need for a TTT to pin down meaning, the need seems
>ph> almost a piece of philosophical nitpicking, since one can get so far -
>ph> in fact, can probably do all of the science - without ever really
>ph> conceding it. In terms of thinking about actual AI work, the difference
>ph> between TT and TTT doesn't really MATTER. And by the way, if one talks
>ph> to people in CS, they often tend to regard the term 'computation' as
>ph> including for example real-time control of a lime kiln.

I don't think you need the TTT to "pin down" meaning. I think you need the structures and processes that make it possible to pass the TTT in order to implement meaning at all. And that definitely includes transduction.

We don't disagree, by the way, on the power of computation to capture and help us understand, explain, predict and build just about anything (be it planes or brains). I just don't think computation alone can either fly or think.

>ph> Heres a question: never mind transducers (I never liked that concept
>ph> anyway), how about proprioception? How much of a sense of pain can be
>ph> accounted for in terms of computations? This is all internal, but I
>ph> bet we need a version of the TTT to ultimately handle it properly, and
>ph> maybe one gets to the physics rather more immediately, since there isn't
>ph> any place to draw the sensed/sensing boundary.

>ph>
>ph> Pat Hayes

As I wrote in my comment on Brian Smith's contribution, this conflating (i) internal/external with respect to a robot's BODY (which is no problem, and may involve lots of "internal" transducers -- for temperature, voltage, etc. -- that are perfectly analog rather than symbolic, despite their internal locus) with (ii) internal/external with respect to the robot's MIND:

(1) What is "in" the mind is certainly inside the body (though "wide intentionalists" tend to forget this); but

(2) what is in the mind is not necessarily symbolic;

(3) what is inside the body in not necessarily symbolic;

(4) what is inside the body is not necessarily in the mind.

The question is not how to "account for" or "handle" proprioception or pain, but how to EMBODY them, how to implement them. And I'm suggesting that you can't implement them AT ALL with computation alone -- not that you can't implement them completely or unambiguously that way, but that you can't implement them AT ALL. (Or, as an intuition pump, you can implement pain or proprioception by computation alone to the same degree that you can implement flying or heating by computation alone.)

Stevan Harnad

Date: Tue, 21 Apr 92 20:48:01 EDT From: "Stevan Harnad"

Below is a contribution to the symbol grounding discussion from Mike Dyer. I will not reply here, because the disagareement between Mike and me has already appeared in print (Dyer 1990, Harnad 1990 in the same issue of JETAI; my apologies for not having the page span for Mike's article at hand).

I will just point out here that Mike seems prepared to believe in some rather radical neurological consequences following from the mere memorization of meaningless symbols. To me this is tantamount to sci-fi. Apart from this, I find that the variants Mike proposes on Searle's Argument seem to miss the point and change the subject.

Stevan Harnad

Harnad, S. (1990) Lost in the hermeneutic hall of mirrors. Invited Commentary on: Michael Dyer: Minds, Machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence 2: 321 - 327.

Date: Tue, 14 Apr 92 23:19:41 PDT From: Dr Michael G Dyer Subject: networks gating networks: minds supervening on minds

Stevan,

It appears that your unfortunate blind acceptance of Searle’s Chinese Room Agument (CRA) keeps leading you astray. In your analysis of Chalmers’s observations you at least correctly grasp that

"So if a mind supervenes on (the right) computations because of their computational properties (rather than because of the physical details of any particular implementation of them), then it must supervene on ALL implementations of those computations."

But then you get derailed with:

"I think Searle’s Chinese Room Argument has successfully pointed out that this will not be so..."

But Searle’s CRA has NOT been "successful". CRA is quite FLAWED, but you don’t seem to entertain any notions concerning how two brains/minds might be intertwined in the same body.

Scenario #1: Suppose the Chinese instructions that Searle follows actually cause Searle to simulate the activation of a complex network of artificial neurons, equal in complexity to the neural network of a human brain (just what’s in those instruction books is never specified, so I can imagine anything I want). We then take those instructions and build a specialized computer "neuro-circuit" that realizes those instructions -- call it C. We then enlarge Searle’s head and install C so that, when "turned on" it takes over Searle’s body. With a remote control device we turn C on and suddenly Searle starts acting like some particular Chinese individual. Once turned on, the Chinese persona requests to maintain control of the body that once was Searle’s.

Scenario #2: We build, on a general purpose multiprocessor, a simulation of the neuro-circuitry of C -- let’s call this Simu-C -- such that SC takes over the body when we flip a switch.

Scenario #3: This one is a bit trickier. We examine carefully Searle’s own neuro-circuitry and we design an artificial neural network -- call it C-supervene -- that gates the circuits of Searle’s brain such that, when C-supervene is turned on, the gating of Searle’s circuitry causes Searle’s own brain circuitry to turn into the Chinese person circuitry. Thus, Searle’s own neurocircuitry is being used (in a rather direct way -- i.e. no symbols) to help create the Chinese personage.

Scenario #4; But now we replace C-supervene with a general multi- processor that runs a simulation (ie. symbol manipulations) that gates Searle’s own neuro-circuitry to produce the Chinese persona.

In scenario #1 there are two distinct sets of neuro-circuits: Searle’s and the Chinese person’s. Whichever one controls the body depends on our switch.

In scenario #2 the Chinese neurocircuitry is replaced by a simulation of that circuitry with a more general purpose hardware.

In scenario #3 the Chinese neurocircuitry actually makes use of Searle’s neurocircuitry to do its computations, but it is the "master" and Searl’es circuitry is the "slave".

In scenario #4, again, the specialized Chinese "neuro-controller" of scenario #3 is replaced by a simulation on a more general purpose hardware.

Finally, we can give the Chinese person (in whichever incarnation above that you want) a set of instructions that allows it to simulate Searle's entire neuro-circuitry, so that, when we flip our switch, it's the Searle persona who gains control of the body. So we can multiple levels of Searles and the Chinese persons simulating each other.

Now, WHICH person should we listen to? When in control of the body, the Searle persona says he's the real person and he has no experience of being the Chinese person. When the Chinese person is in control, this person claims to have no experience of being Searle (and so uses Searle's own argument against Searle).

Now, it is still possible that there is more to having a mind than having the right sort of computations, but Searle has NOT given any kind of refutation with his CRA. And your own "grounding" argument is insufficient also since it leads to either one of two absurd situations:

either you have to claim that (a) remove the eyes and the mind goes. or (b) the entire brain has to count as the eyes, so that you get to remove the entire brain whenever anyone requests that you remove the eyes.

On the other side, if we accept Chalmers's position (and the working hypothesis of the Physical Symbol System Hypothesis of AI), then we have the advantage of being able to compare minds by observing how similar their computations are (at some level of abstraction) and we can develop, ultimately, a non-chauvinistic theory of mind (e.g. alien intelligences in different substrata).

Notice, connectionism fits within the PSSH (because every neuron and synapse and dendritic compartment, etc. can be represented by a symbol and its behavior modeled by symbol manipulation).

Chalmers (and AI researchers) may be wrong, but Searle (and you) have not given any kind of ironclad arguments that they are.

So let's hear something from you OTHER than your overly and oft used argument of the form:

"X..., but we know Searle's CRA is right, so X can't be..."

(There are large numbers of AI researchers out here who are not convinced even one whit by Searle's CRA or your "out of sight out of mind" argument. So you and Searle need to come up with something new. Searle's INABILITY to "feel" what it's like to "be" the Chinese person he's bringing into existence IS TO BE EXPECTED. Nor do we expect the the Lisp interpreter to know what it's like to be the particular expert system that "supervenes" upon it.)

-- Michael Dyer

Date: Wed, 22 Apr 92 13:06:13 EDT From: "Stevan Harnad"

Date: Wed, 22 Apr 92 12:00:06 +0200 From: tim@arti1.vub.ac.be (Tim Smithers)

I have two responses to Mike Dyer's recent contribution to your symbol grounding discussion.

First, the actual practice of (symbol processing) AI research makes it very difficult to talk about the Physical Symbol System Hypothesis (PSSH) of Newell and Simon as being "a working hypothesis". It is much more a widely accepted and unquestioned dogma than it is a hypothesis. For it to be a hypothesis, in the normal scientific sense (symbol processing) AI research would need to be conducting a broad experimental investigation of its validity (or otherwise). Very little, if any, research is either presented as, or can be properly understood to be, a contribution to such a research programme. Talk of constructing theories of mind (even non-chauvinistic ones) given the current practice thus seems to me to be unsupportable.

Second, even if there are lots of AI researchers out there who are not convinced by Searle's and your arguments, does Mike Dyer think that they are going to be convinced by such in principle possibly true but in practice right out of the window aguments as he offers? I think all you guys should spend some time trying to build real robots (not simulations!) that reliably do real things (even very simple things) in the real world. This might introduce some reality constraints on this whole discussion from which I think it would benefit enormously--it might 'ground' the discussion, if you don't mind my misusing your term.

Tim Smithers

From: Pat Hayes Date: Sun, 19 Apr 92 15:08:06 MDT To: Brian C Smith

Brian-

As usual, you say things clearly. But I can't resist a few comments. I follow your numbering.

1. I agree wholeheartedly with your observations about the limitations of universality. But not with your conclusions. Searle talks about the distinction between a model and the real thing, but the moral of the classical work on universality (and of CS practice - not just in Silicon Valley, by the way!) is exactly that a computational simulation of a computation IS a computation. Thus, a LISP interpreter running LISP really is running LISP: it's no less really computation than if one had hardware devoted to the task.

That is I think a crucial insight, perhaps the central one of CS, and one which was historically very surprising. That's why computers work, why we can run LISP and Word-4 on the same machine. To put that aside as a red herring is to simply regard computers as fast switching devices. It carries possible implications for biology, for example it suggests an account of why evolution produced so much cognition so quickly. While this idea and its implications is a minefield, I think it's one we need to be treading through, and definitely not any kind of colored fish.

2. [on original intentionality] Oh, I agree! This deserves expanding. Many of the Searlean writers have taken it as somehow axiomatic that human thinking just has this vital property of being meaningful, something that only human, or maybe organic, thinking has been observed to possess. Whatever this is, it isn't anything like a color or a mass that human thought has.

I have a number of beliefs about ancient Rome. How are these thoughts connected to the Rome of 2000 years ago? The answer is probably very complicated, involving texts written by others and translated from language to another, to historians best attempts to reconstruct facts from flimsy evidence, and so forth. The connection between me and Caesar goes through an entire society, indeed a historical chain of societies. My thoughts about Julius Caesar are not somehow intrinsically about him by virtue of their being in my head; but they are in fact about him. But I can't see any reason why a machine could not have almost the same (very complicated) relationship to him that I have, whatever it is, since it is mediated almost entirely by language.

5. [on types] I agree that there is a danger of imposing too much structure on the world, but have a couple of caveats. First, what makes you think that this will ever be completely avoidable? We must use some concepts to build our theories from, and to try to avoid it altogether is not just tough sledding but I think trying to walk into the snow naked (and you know what happened to him.) We have to be conscious of what we are doing, but we must use something, surely. And second, I don't think you are right to dismiss set theory so quickly. Set theory doesn't preparse the world: it only insists that some parsing is made. Others have tried to get by with less, and indeed several other alternatives are available, as well as several alternative foundations. But again, you have to stand somewhere, and these carefully developed, thoroughly tested and well-understood patches of intellectual ground have a lot to recommend them. And I don't think one does have to have the types all specified in advance.

Take maps for example. One can give an account of how maps relate to terrain which assumes that maps have some kind of parsing into meaningful symbols (towns, roads, etc) which denote...well, THINGS in the territory, and talk about a certain class of (spatial) relations between these things which is reflected by a (more-or-less) homomorphic image of them holding between the syntactic objects in the map. Now, there are all sorts of complexities, but the essential idea seems coherent and correct. But notice it has to assume that the terrain can be somehow divided into pieces which are denoted by the map's symbols (or better, that appropriate structures can be found in the terrain). You might object at this point that this is exactly what you are complaining about, but if so I would claim not. Here, the theorist is only assuming that SOME parsing of the terrain can be made: it was the maker of the map who parsed his territories, and the semantic account has to reflect this ontological perspective. So what the theorist needs is some way of describing these ontological complexities which imposes a minimum of structure of its own, and structure which is well-understood so that we can consciously allow for possible distortions it might introduce. And that is just what is provided by the idea of a set. To be sure, there are some artifacts produced by, for example, the conventional extensional representation of functions. But these are immediately recognisable when they occur and have known ways around them.

I once tried to focus on the hardest case I could find for a set-theoretic account, which was the idea of a piece of liquid (since what set theory does seem to assume is the notion of an individual thing conceptually distinct from others, and liquids are very intermergable). And to my surprise, the little intellectual discipline imposed by the use of sets actually clarified the semantic task: it was as though the sets imposed distinctions which were a useful ontological discovery. I have since come to think not that a particular set of types is fixed in advance, but that what does seem to be fixed in us, in our way of thinking, is a propensity to individuate. The world is a continuum, but we see it and think of it as made of things, maybe overlapping in complex ways, but conceptually separate entities that we can name and classify.

This may not be the place to indulge in this discussion, since it is getting away from what computation is, but you asked for things onto the table...

Pat Hayes

From: Pat Hayes To: Stevan Harnad (harnad@princetone.edu) Date: Tue, 21 Apr 92 17:56:09 MDT

>
>sh> As a first pass at "formal," how about: A symbol system
>
>sh> consists of a set of objects (elementary symbols and composite
>
>sh> symbols) plus rules for manipulating the symbols. The rules
>
>sh> operate only on the physical shapes of the symbols, not their
>
>sh> meanings (and the shapes of the elementary symbols are
>
>sh> arbitrary), yet the symbols are systematically interpretable as
>
>sh> meaning something. The rules for manipulating the symbols on
>
>sh> the basis of their shapes are called "syntactic" or "formal"
>
>sh> rules.

Heres an example adapted from one of Brian’s. Take a set of rules which encode (a formal system for) arithmetic, together with a formal predicate ’lengthof’, and the rules

lengthof(’0’) -> 1 lengthof(n<>x) -> lengthof(n) + lengthof(x)

Now, these rules make ’lengthof(n)’ evaluate to (a numeral which means) the number of digits in the formal representation of n: ie, the length of that numeral in digits. Notice this is the ACTUAL length of that piece of syntax. Now, is this ’formal’? It is according to your definition, and perhaps you are happy with that, but it has some marks which successfully refer to physical properties of part of the world.

>
>sh> "intuition pump" is not a pejorative, if it pumps true.

It IS a pejorative if the pump is claimed to be a conclusive argument from obvious assumptions. My intuition tells me clearly that when I debug a piece of code by pretending to be an interpreter and running through it ’doing’ what it ’tells’ me to do, that the program is not being run, and certainly not run on, or by, me. So we are left with your intuition vs. my intuition, and they apparently disagree.

>
>sh> I will be happy to consider the implications of the fact that
>
>sh> Searle, doing everything the computer does, does not count as a
>
>sh> valid implementation of the same computer program -- as soon as
>
>sh> you specify and argue for what you mean by implementation and
>
>sh> why Searle's would not qualify. Until then, I don't see why
>
>sh> EVERY system that processes the same symbols, follows the same
>
>sh> (syntactic) rules and steps through the same states doesn't
>
>sh> qualify as a valid implementation of the same program.

The key is that Searle-in-the-room is not doing everything the computer 'does', and is not going through the same series of states. For example, suppose the program code at some point calls for the addition of two integers. Somewhere in a computer running this program, a piece of machinery is put into a state where a register is CAUSED to contain a numeral representing the sum of two others. This doesn't happen in my head when I work out, say, 3340 plus 2786, unless I am in some kind of strange arithmetical coma. If Searle-in-the-room really was going through the states of an implementation of a chinese-speaking personality, then my intuition, pumped as hard as you like, says that that Chinese understanding is taking place. And I haven't yet heard an argument that shows me wrong.

>
>sh> I think most of the brain is preserving sensory signals in
>
>sh> various degrees of analog form (so we would probably do well to
>
>sh> learn from this).

While we should have awe for what Nature has wrought, we also must keep our wits about us. The reason elephants have big brains is that they have a lot of skin sending in signals which need processing, and the neurons come in at a certain density per square inch. This is evolution's solution to the bandwidth problem: duplication. Similarly, that the motor and sensory cortex use 'analogical' mappings of bodily location is probably more due to the fact that this fits very nicely with the way the information is piped into the processor, where location is encoded by neuroanatomy, than by any profound issue about symbolic vs. analog. It has some nice features, indeed, such as localisation of the effects of damage: but we are now in the language of computer engineering.

>
>sh> In fact, I think it's as likely that a mind is mostly symbolic,
>
>sh> with just a thin analog layer mediating input and output to the
>

>sh> world, as that a plane or a furnace are mostly symbolic, with a
>
>sh> thin analog layer mediating input and output.

You are exhibiting here what I might call Searleanism. Of course a furnace is not symbolic. But hold on: thats the point, right? Furnaces just operate in the physical world, but minds (and computers) do in fact react to symbols: they do what you tell them, or argue with you, or whatever: but they respond to syntax and meaning, unlike furnaces and aircraft. That's what needs explaining. If you are going to lump furnaces and minds together, you are somehow missing the point that drives this entire enterprise. (Aircraft are actually a borderline case, since they do react to the meanings of symbols input to them, exactly where they have computers as part of them.)

>
>sh> ... I'm suggesting that you can't implement them AT ALL with
>
>sh> computation alone -- not that you can't implement them
>
>sh> completely or unambiguously that way, but that you can't
>
>sh> implement them AT ALL.

I agree, because with what you mean by computation, I couldn't even run Wordstar with computation ALONE. I need a computer.

>
>sh> (Or, as an intuition pump, you can implement pain or
>
>sh> proprioception by computation alone to the same degree that you
>
>sh> can implement flying or heating by computation alone.)

I bet computational ideas will be centrally involved in a successful understanding of pain and proprioception, probably completely irrelevant to understanding lime chemistry, but important in reasonably exotic flying.

But now we are just beating our chests at one another. Like I said, its only a pump, not an argument.

Pat Hayes

Date: Wed, 22 Apr 92 15:12:37 EDT From: "Stevan Harnad"

ON SYNTHETIC MINDS AND GROUNDED "ABOUTNESS"

Pat Hayes (phayes@nmsu.edu) wrote:

>ph> Many of the Searlean writers have taken it as somehow axiomatic that
>ph> human thinking just has this vital property of being meaningful,
>ph> something that only human, or maybe organic, thinking has been observed
>ph> to possess. Whatever this is, it isn't anything like a color or a mass
>ph> that human thought has.

Not sure who "Searlean" writers are, but this writer certainly does not claim that only organic thinking is possible (e.g., I am working toward TTT-scale grounded robots). No one has given any reason to believe that synthetic minds can't be built. Only one candidate class of synthetic minds has been ruled out by Searle's argument, and that is purely computational ones: stand-alone computers that are merely running the right software, i.e., any and all implementations of symbol systems that allegedly think purely because they are implementations of the right symbol system: the symbol system on which the mind "supervenes" (with the implementational particulars being inessential, hence irrelevant).

But there are plenty of other candidates: Nonsymbolic systems (like transducers and other analog devices), hybrid nonsymbolic/symbolic systems (like grounded robots), and even implemented symbol systems in which it is claimed that specific particulars of the implementation ARE essential to their having a mind (Searle's argument can say nothing against those, because he couldn't BE the system unless its implementational details were irrelevant!).

>ph> I have a number of beliefs about ancient Rome. How are these thoughts
>ph> connected to the Rome of 2000 years ago? The answer is probably very
>ph> complicated... a historical chain... My thoughts about Julius Caesar
>ph> are not somehow intrinsically about him by virtue of their being in my
>ph> head; but they are in fact about him. But I can't see any reason why a
>ph> machine could not have almost the same (very complicated) relationship
>ph> to him that I have, whatever it is, since it is mediated almost
>ph> entirely by language.

I see no reason why a grounded (TTT-indistinguishable) robot's thoughts would not be just as grounded in the objects they are systematically interpretable as being about as my own thoughts are. I diverge from Searle and many other post-Brentano/Fregean philosophers in denying completely that there are two independent mind/body problems, one being the problem of consciousness (qualia) and the other being the problem of "aboutness" (intentionality). In a nutshell, there would be no problem of thoughts having or not having real "aboutness" if there were not something it was like to think (qualia). The reason the symbols in a computer are not "about" anything is because there's nobody home in there, consciously thinking thoughts!

[This is what is behind the force of Searle's simple reminder that he would surely be able to state, with complete truthfulness, that he had no idea what he was talking about when he "spoke" Chinese purely in virtue of memorizing and executing the very same syntactic symbol-manipulation operations that are performed by the TT-passing computer. We each know exactly what it is LIKE to understand English, what it is LIKE to mean what we mean when we speak English, what it is LIKE for our words to be about what they are about; no such thing would be true for Searle, in Chinese, under those conditions. Hence the fact that the Chinese input and output was nevertheless systematically (TT) interpretable AS IF it were about something would merely show that that "aboutness" was not "intrinsic," but derivative, in exactly the same sense that it would be derivative in the case of the symbols in an inert book, in which there is likewise nobody home.]

On the other hand, there is still the POSSIBILITY that grounded TTT-scale (performance indistinguishable) robots or even grounded TTTT-scale (neurally indistinguishable) robots fail to have anybody home in them either. Now that IS the (one, true) mind/body problem, but we should be ready to plead no contest on that one (because the TTT and the TTTT take us to the limits of empiricism in explaining the mind).

>ph> I... think not that a particular set of types is fixed in advance, but
>ph> that what does seem to be fixed in us, in our way of thinking, is a
>ph> propensity to individuate. The world is a continuum, but we see it and
>ph> think of it as made of things, maybe overlapping in complex ways, but
>ph> conceptually separate entities that we can name and classify.
>ph>
>ph> Pat Hayes

Hence we should investigate and model the structures and processes underlying our capacity to categorize inputs (beginning with sensory projections). Those structures and processes will turn out to be largely nonsymbolic, but perhaps symbols can be grounded in the capacity those nonsymbolic structures and processes give us to pick out the objects they are about.

Stevan Harnad

Harnad, S., Hanson, S.J. & Lubin, J. (1991) Categorical Perception and the Evolution of Supervised Learning in Neural Nets. In: Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology (DW Powers & L Reeker, Eds.) pp. 65-74. Presented at Symposium on Symbol Grounding: Problems and Practice, Stanford University, March 1991; also reprinted as Document D91-09, Deutsches Forschungszentrum fur Kuenstliche Intelligenz GmbH Kaiserslautern FRG.

Andrews, J., Livingston, K., Harnad, S. & Fischer, U. (1992) Learned Categorical Perception in Human Subjects: Implications for Symbol Grounding. Proceedings of Annual Meeting of Cognitive Science Society (submitted)

Harnad, S. Hanson, S.J. & Lubin, J. (1992) Learned Categorical Perception in Neural Nets: Implications for Symbol Grounding. Proceedings of Annual Meeting of Cognitive Science Society (submitted)

From: Pat Hayes To: Stevan Harnad (harnad@princeton.edu) Date: Tue, 21 Apr 92 17:56:09 MDT

>
>sh> As a first pass at "formal," how about: A symbol system
>
>sh> consists of a set of objects (elementary symbols and composite
>
>sh> symbols) plus rules for manipulating the symbols. The rules
>
>sh> operate only on the physical shapes of the symbols, not their

>
>sh> meanings (and the shapes of the elementary symbols are
>
>sh> arbitrary), yet the symbols are systematically interpretable as
>
>sh> meaning something. The rules for manipulating the symbols on
>
>sh> the basis of their shapes are called "syntactic" or "formal"
>
>sh> rules.

Heres an example adapted from one of Brian’s. Take a set of rules which encode (a formal system for) arithmetic, together with a formal predicate ’lengthof’, and the rules

lengthof(’0’) -> 1 lengthof(n<>x) -> lengthof(n) + lengthof(x)

Now, these rules make ’lengthof(n)’ evaluate to (a numeral which means) the number of digits in the formal representation of n: ie, the length of that numeral in digits. Notice this is the ACTUAL length of that piece of syntax. Now, is this ’formal’? It is according to your definition, and perhaps you are happy with that, but it has some marks which successfully refer to physical properties of part of the world.

>
>sh> "intuition pump" is not a pejorative, if it pumps true.

It IS a pejorative if the pump is claimed to be a conclusive argument from obvious assumptions. My intuition tells me clearly that when I debug a piece of code by pretending to be an interpreter and running through it ’doing’ what it ’tells’ me to do, that the program is not being run, and certainly not run on, or by, me. So we are left with your intuition vs. my intuition, and they apparently disagree.

>
>sh> I will be happy to consider the implications of the fact that
>
>sh> Searle, doing everything the computer does, does not count as a
>
>sh> valid implementation of the same computer program -- as soon as
>
>sh> you specify and argue for what you mean by implementation and
>
>sh> why Searle’s would not qualify. Until then, I don’t see why
>
>sh> EVERY system that processes the same symbols, follows the same
>
>sh> (syntactic) rules and steps through the same states doesn’t
>
>sh> qualify as a valid implementation of the same program.

The key is that Searle-in-the-room is not doing everything the computer ’does’, and is not going through the same series of states. For example, suppose the program code at some point calls for the addition of two integers. Somewhere in a computer running this program, a piece of machinery is put into a state where a register is CAUSED to contain a numeral representing the sum of two others. This doesn’t happen in my head when I work out, say, 3340 plus 2786, unless I am in some kind of strange arithmetical coma. If Searle-in-the-room really was going through the states of an implementation of a chinese-speaking personality, then my intuition, pumped as hard as you like, says that that Chinese understanding is taking place. And I haven’t yet heard an argument that shows me wrong.

>
\>sh> I think most of the brain is preserving sensory signals in
>
\>sh> various degrees of analog form (so we would probably do well to
>
\>sh> learn from this).

While we should have awe for what Nature has wrought, we also must keep our wits about us. The reason elephants have big brains is that they have a lot of skin sending in signals which need processing, and the neurons come in at a certain density per square inch. This is evolution’s solution to the bandwidth problem: duplication. Similarly, that the motor and sensory cortex use ’analogical’ mappings of bodily location is probably more due to the fact that this fits very nicely with the way the information is piped into the processor, where location is encoded by neuroanatomy, than by any profound issue about symbolic vs. analog. It has some nice features, indeed, such as localisation of the effects of damage: but we are now in the language of computer engineering.

>
\>sh> In fact, I think it’s as likely that a mind is mostly symbolic,
>
\>sh> with just a thin analog layer mediating input and output to the
>
\>sh> world, as that a plane or a furnace are mostly symbolic, with a
>
\>sh> thin analog layer mediating input and output.

You are exhibiting here what I might call Searleanism. Of course a furnace is not symbolic. But hold on: thats the point, right? Furnaces just operate in the physical world, but minds (and computers) do in fact react to symbols: they do what you tell them, or argue with you, or whatever: but they respond to syntax and meaning, unlike furnaces and aircraft. That’s what needs explaining. If you are going to lump furnaces and minds together, you are somehow missing the point that drives this entire enterprise. (Aircraft are actually a borderline case, since they do react to the meanings of symbols input to them, exactly where they have computers as part of them.)

>
\>sh> ... I’m suggesting that you can’t implement them AT ALL with
>
\>sh> computation alone -- not that you can’t implement them
>

>sh> completely or unambiguously that way, but that you can't
>
>sh> implement them AT ALL.

I agree, because with what you mean by computation, I couldn't even run Wordstar with computation ALONE. I need a computer.

>
>sh> (Or, as an intuition pump, you can implement pain or
>
>sh> proprioception by computation alone to the same degree that you
>
>sh> can implement flying or heating by computation alone.)

I bet computational ideas will be centrally involved in a successful understanding of pain and proprioception, probably completely irrelevant to understanding lime chemistry, but important in reasonably exotic flying.

But now we are just beating our chests at one another. Like I said, its only a pump, not an argument.

Pat Hayes

Date: Wed, 22 Apr 92 17:39:44 EDT From: "Stevan Harnad"

Pat Hayes wrote:

>ph> Here's an example adapted from one of Brian [Smith's]. Take a set
>ph> of rules which encode (a formal system for) arithmetic, together with
>ph> a formal predicate 'lengthof', and the rules
>ph>
>ph> lengthof('0') -> 1
>ph> lengthof(n<>x) -> lengthof(n) + lengthof(x)
>ph>
>ph> Now, these rules make 'lengthof(n)' evaluate to (a numeral which means)
>ph> the number of digits in the formal representation of n: ie, the length
>ph> of that numeral in digits. Notice this is the ACTUAL length of that
>ph> piece of syntax. Now, is this 'formal'? It is according to your
>ph> definition, and perhaps you are happy with that, but it has some marks
>ph> which successfully refer to physical properties of part of the world.

It is a very interesting and useful feature of symbol systems that some can be formulated so as to be INTERPRETABLE as referring to themselves (as in the sentence "this sentence has five words") or to physical properties (especially numerical ones) of other symbols and symbol strings within the same system. Symbol systems that go on to USE the nonarbitrary analog properties of their symbol tokens as data are special in certain respects (as in "numeric" versus "symbolic" computation) and may cast just a bit more light on the dynamics of dedicated hybrid symbolic/analog systems, and perhaps even on symbol grounding. I don't know.

But note that in your example above, even though the computation yields a symbol that is interpretable as the number of symbols in the string, this is in principle no different from a computation that yields a symbol that is interpretable as the number of planets in the solar system. It is just a systematic correspondence (and hence interpretable as such). But "interpretable as meaning X" (as in the case of a book, interpretable by a thinking mind) is not the same as "meaning X" (as in the case of thoughts, in a mind). Failing to distinguish the two seems to be another instance of conflating physical inner/outer and mental inner/outer, as discussed earlier.

>ph> My intuition tells me clearly that when I debug a piece of code by
>ph> pretending to be an interpreter and running through it "doing" what it
>ph> "tells" me to do, that the program is not being run, and certainly not
>ph> run on, or by, me. So we are left with your intuition vs. my intuition,
>ph> and they apparently disagree.

But isn't the real question whether there is any relevant difference between what you think is a "real" implementation by a machine and what you think is a "pseudo-implementation" by a person? Certainly the computer is not stepping through the states consciously and deliberately, as you are. But is there anything else that's different? If we speak only of the "motions gone through" and their I/O conditions in the two cases, they are exactly the same. In the case of the machine, the motions are mechanical; no choice is involved. In the case of the person, their elective. But so what? Even apart from the vexed questions associated with free will and causality, what is there about taking IDENTICAL motions under identical I/O conditions and making their causal basis mindless and mechanical that could possibly effect a transition INTO the mental (rather than OUT of it, which is the much more obvious feature of the transition from the human implementation to the machine one)?

>ph> The key is that Searle-in-the-room is not doing everything the computer
>ph> "does," and is not going through the same series of states. For
>ph> example, suppose the program code at some point calls for the addition
>ph> of two integers. Somewhere in a computer running this program, a piece
>ph> of machinery is put into a state where a register is CAUSED to contain
>ph> a numeral representing the sum of two others. This doesn't happen in my
>ph> head when I work out, say, 3340 plus 2786, unless I am in some kind of
>ph> strange arithmetical coma. If Searle-in-the-room really was going
>ph> through the states of an implementation of a chinese-speaking
>ph> personality, then my intuition, pumped as hard as you like, says that
>ph> that Chinese understanding is taking place. And I haven't yet heard an
>ph> argument that shows me wrong.

It's always useful, in this sort of hermeneutic puzzle, to de-interpret and reduce things to gibberish as much as possible: Suppose the computer was doing all the requisite summation in binary, and you were too, and all it did, and all you did, was compare zero's and one's and erase and carry, just like a Turing Machine. Is it still so obvious that you're not doing everything the computer is doing? If anything, the computer is doing less than you rather than more (because it has no choice in the matter). Why should I interpret less as more?

>ph> that the motor and sensory cortex use 'analogical' mappings of bodily
>ph> location is probably more due to the fact that this fits very nicely
>ph> with the way the information is piped into the processor, where

>ph> location is encoded by neuroanatomy, than by any profound issue about
>ph> symbolic vs. analog. It has some nice features, indeed, such as
>ph> localisation of the effects of damage: but we are now in the language
>ph> of computer engineering.

I too am thinking of this only as (reverse bio-)engineering. But brains can do so much more than any machine we've yet engineered, and they seem to do so much of it in analog. It seems that this might be a useful cue to take, but maybe not. It's an empirical question.

>ph> Of course a furnace is not symbolic. But hold on: that's the point,
>ph> right? Furnaces just operate in the physical world, but minds (and
>ph> computers) do in fact react to symbols: they do what you tell them, or
>ph> argue with you, or whatever: but they respond to syntax and meaning,
>ph> unlike furnaces and aircraft. That's what needs explaining. If you are
>ph> going to lump furnaces and minds together, you are somehow missing the
>ph> point that drives this entire enterprise. (Aircraft are actually a
>ph> borderline case, since they do react to the meanings of symbols input
>ph> to them, exactly where they have computers as part of them.)

I don't think I'm missing the point. Computation has been able to generate some very fancy and flexible performance -- certainly fancier and more flexible than that of a furnace or plane (except "smart," computer-aided planes, as you indicate). Computation also seems to resemble thought in its syntactic structure. It was accordingly quite reasonable to hypothesize that thinking -- that unobservable process going on in our heads -- might actually be a form of computation. But here we are discussing reasons why, despite promising initial appearances, that hypothesis is turning out to be WRONG, and what is going on in our heads is something else, not computation (or not just computation).

By the way, minds and computers may both respond to syntax, but only minds respond to meaning. Computers are merely INTERPRETABLE as if they responded to meaning...

>ph> with what you mean by computation, I couldn't even run
>ph> Wordstar with computation ALONE. I need a computer.

Pat, you know I stipulated that the computation had to be physically implemented; I just stressed that the particulars of the implementation (apart from the fact that they stepped through the right states with the right I/O) were irrelevant.

>ph> I bet computational ideas will be centrally involved in a successful
>ph> understanding of pain and proprioception, probably completely
>ph> irrelevant to understanding lime chemistry, but important in reasonably
>ph> exotic flying.
>ph>
>ph> Pat Hayes

And I bet a lot of the essential features of pain and proprioception will be in the analog properties of the hardware that implements it, which will be more like exotic chemistry.

Stevan Harnad

Date: Mon, 20 Apr 92 03:23:41 -0400 From: yee@envy.cs.umass.edu Subject: Don't talk about "computers"

I would like to follow up on some of Brian Smith's recent comments regarding universal computations and formal/non-formal symbol processing. I propose that we try to avoid using the terms "computer" and "program" because they are misleading with regard to questions of the computability of mind. For "computer" I would use "Turing machine" and I generally would not discuss programs because they are just descriptions of TM's.

The things we usually refer to as "computers" are physical instantiations of universal Turing machines (UTM's), a particular subclass of Turing machine (TM). Unfortunately, philosophical discussions about computers (UTM's) generally carry an implicit extension to all TM's. Presumably, this occurs because UTM's are "universal." But as Brian indicated, UTM universality refers to a very special type of *weak equivalence* (Pylyshyn, 1984) between TM's and UTM's. Universality merely means partial I/O equivalence. This is insufficient for many discussions about the computability of mind---e.g., the Chinese Room---because such discussions consider, not only I/O behavior, but also *how* the behavior is achieved, and UTM's are far from "typical" in their manner of computation. In particular, although UTM's process certain input symbols purely formally, not all TM's need behave this way.

To review briefly, any program P describes a Turing machine Tp that maps inputs x to outputs y (as shown below in A). Any UTM U (shown in B) is special in that its inputs z are composites of a program P and a nominal-input x', i.e., z=(P,x').

+----+ x'-+ +---+ x -->| Tp |--> y +- z ->| U |--> y +----+ P -+ +---+

(A) a TM. (B) a UTM.

Formal symbol processing of nominal-inputs by UTM's is a special consequence of their being given input programs. A UTM U can always produce output y by processing nominal-input x'=x purely formally because P completely controls the processing of x', independently of U. That is, U's computation on z simply instantiates Tp's computation x --> y.

Clearly, U's formal treatment of x' does not imply that Tp's processing of x is necessarily formal. Such a conclusion would require a special proof. For all we know, Tp might associate x with internally stored information and produce output y accordingly. One might try to show that all TM's are restricted to formal symbol processing, but this would not follow automatically from the fact that UTM's can get away with formally processing (a portion of) their inputs. (Actually, in a paper cited below I argue that, in general, TM's can process symbols non-formally.)

+-----{ CR }-------+ | | x'-------+ +---+ | | +- z ->| U |----> y | p -+ +---+ | +------------------+

(C) a UTM viewed as the CR.

The implications of the TM/UTM distinction for the Chinese Room (CR) argument are straightforward. The person in the CR is a UTM U that is given a program P (the rules). (Note that "memorizing" the rules does not change U into Tp. Any set of rules could be memorized, and the memorized rules remain an input to U.) To answer the question of *how* the Chinese symbols x' are being processed inside the room, one must consider what *Tp* is doing to the symbols. Considering only U's activity is useless because U is computing z=(P,x')--> y. Thus, without specific knowledge of the rules P, one simply cannot answer the question of whether the Chinese input symbols are being understood in the CR or are only being formally manipulated. Both possibilities remain open (unless, of course, one advocates the Turing Test for understanding, but that is an independent argument).

In general, if one wants to know how a program P really operates, then one should only consider the corresponding Turing machine Tp. If you build a UTM U, give it P, and then look at what U is doing, you will be looking in the wrong place. Eliminate the middleman, and build Tp directly.

Finally, one concludes that an assertion such as

Every computer has property X. (1)

is generally ambiguous and should be replaced by either

Every TM has property X. (2a) or Every UTM has property X. (2b)

Clearly, (2a) implies (2b), but not conversely. At best, the Chinese Room argument shows that UTM computations are not good candidates for minds. However, there remain plenty of non-universal TM computations, and---absent any proof to the contrary---some of them might be minds. To find out, one should forget about computers and think about instantiated programs. If one's real interest is the entire class of TM's, then it is dangerous to form intuitions and conclusions revolving around the special properties of UTM's.

Many debates about computers and minds pit critics of purely formal symbol processing (which UTM's perform) against proponents of computation (which all TM's perform). A failure to clearly maintain the TM/UTM distinction means that, not surprisingly, discussants often appear to talk past each other. Nevertheless, it remains entirely consistent to believe (the *correct* :-) portions of both the "semantophiles'" and the "computationalists'" arguments. That is, intentionality, symbol-grounding, meaning, etc. (of the type desired by Searle, Harnad, Penrose and others) is necessary for (human-like) minds, and such semantics is Turing-computable.

Richard Yee

@Book{Pylyshyn:84, author = "Pylyshyn, Z. W.", title = "Computation and Cognition: Toward a Foundation for Cognitive Science", publisher = "Bradford Books/MIT Press", year = "1984", address = "Cambridge, MA",

@Unpublished{Yee:rcssp, author = "Yee, Richard", title = "Real Computers and Semantic Symbol Processing", note = "Dept.\ of Computer Science, Univ. of Massachusetts, Amherst, MA 01003. E-mail: yee@cs.umass.edu", year = "1991", month = "March"

Date: Wed, 22 Apr 92 19:14:07 EDT From: "Stevan Harnad"

SO WHAT IS COMPUTATION?

In his comment entitled "Don't talk about computers," Richard Yee (yee@envy.cs.umass.edu) wrote:

>ry> as Brian [Smith] indicated, UTM universality refers to a very special
>ry> type of *weak equivalence* (Pylyshyn, 1984) between TM's and UTM's.
>ry> Universality merely means partial I/O equivalence. This is insufficient
>ry> for many discussions about the computability of mind---e.g., the
>ry> Chinese Room---because such discussions consider, not only I/O
>ry> behavior, but also *how* the behavior is achieved, and UTM's are far
>ry> from "typical" in their manner of computation. In particular, although
>ry> UTM's process certain input symbols purely formally, not all TM's need
>ry> behave this way.

Much of Yee's comment is based an a distinction between formal and nonformal "computation," whereas my arguments are based completely on computation as formal symbol manipulation. We will need many examples of what nonformal computation is, plus a clear delineation of what is NOT nonformal computation, if this is to help us with either the question of what is and is not a computer (or computation) or the question of whether or not mental processes are computational and whether or not computers can have minds. (It would also seem hard to pose these questions without talking about computers, as Yee enjoins us!)

>ry> The implications of the TM/UTM distinction for the Chinese Room (CR)
>ry> argument are straightforward. The person in the CR is a UTM U that is
>ry> given a program P (the rules). (Note that "memorizing" the rules does
>ry> not change U into Tp. Any set of rules could be memorized, and the
>ry> memorized rules remain an input to U.) To answer the question of *how*
>ry> the Chinese symbols x' are being processed inside the room, one must
>ry> consider what *Tp* is doing to the symbols. Considering only U's
>ry> activity is useless because U is computing z=(P,x')--> y. Thus, without
>ry> specific knowledge of the rules P, one simply cannot answer the
>ry> question of whether the Chinese input symbols are being understood in
>ry> the CR or are only being formally manipulated. Both possibilities
>ry> remain open (unless, of course, one advocates the Turing Test for
>ry> understanding, but that is an independent argument).

The Turing Test has been intimately involved in Searle's Argument from the beginning. The Argument is directed against a position Searle dubbed "Strong AI," according to which a computer program that could pass the Turing Test (in Chinese) would understand (Chinese) no matter how it was implemented. Searle simply points out to us that if he himself implemented the program (by memorizing the symbols and symbol manipulation rules) he would not understand Chinese, hence neither would any computer that implemented the same program. So much for the Turing Test and the computationality of understanding.

The only thing that is critical for Searle's argument is that he be able to DO with the input and output exactly the same (RELEVANT) things the computer does. The implementational details are irrelevant; only the program is relevant. And the TT is simply an I/O criterion.

Now I have no idea what YOU are imagining the computer to be doing; in particular, what would it be doing if it were doing "nonformal computation"? If it would be doing something that was not implementation-independent, then you've simply changed the subject (and then even a transducer would be immune to Searle's argument). If it IS doing something implementation-independent, but not "formal," then again, what is it, and can Searle do it or not?

>ry> At best, the Chinese Room argument shows that UTM computations are not
>ry> good candidates for minds. However, there remain plenty of
>ry> non-universal TM computations, and---absent any proof to the
>ry> contrary---some of them might be minds. To find out, one should forget
>ry> about computers and think about instantiated programs. If one's real
>ry> interest is the entire class of TM's, then it is dangerous to form
>ry> intuitions and conclusions revolving around the special properties of
>ry> UTM's.

This won't do at all, because for all I know, I can think of an airplane or a planetary system as an "instantiated program" on a "non-universal TM," and that would make the question of what computers/computation can/cannot do pretty empty. Please give examples of what are and are not "non-universal TM computations" and a principled explanation of why they are or are not.

>ry> Many debates about computers and minds pit critics of purely formal
>ry> symbol processing (which UTM's perform) against proponents of
>ry> computation (which all TM's perform)... intentionality,
>ry> symbol-grounding, meaning, etc. (of the type desired by Searle, Harnad,
>ry> Penrose and others) is necessary for (human-like) minds, and such
>ry> semantics is Turing-computable.
>ry>
>ry> Richard Yee

One cannot make coherent sense of this until the question "What is computation?", as posed in the header to this discussion, is answered. Please reply in ordinary language before turning again to technical formalisms, because this first pass at formalism has merely bypassed the substantive questions that have been raised.

Stevan Harnad

Date: Thu, 23 Apr 92 17:12:30 EDT From: "Stevan Harnad"

SEARLE'S PERISCOPE

>ph> From: Pat Hayes
>ph> Date: Wed, 22 Apr 92 15:30:28 MDT
>ph> To: tim@arti1.vub.ac.be (Tim Smithers)
>ph> Subject: Re: Smithers on Dyer on the physical symbol hypothesis (PSH)
>ph>
>ph> Dear Tim Smithers,
>ph>
>ph> First, the PSH is as much a hypothesis as, say, the hypothesis of
>ph> continental drift. Nobody could observe continental drift or
>ph> conduct a 'broad experimental investigation' of its validity.
>ph> It is a general idea which makes sense of a large number of
>ph> observations and provides a framework within which many empirical
>ph> results can be fitted. Most of the hypotheses of science are like
>ph> this: they aren't tested by little well-designed experiments, and
>ph> indeed couldn't be. There are whole areas of investigation,
>ph> such as cosmology, which couldn't be done in this simplistic
>ph> textbook way of (idea->design experiment->test->next idea),
>ph> and whole methodologies, such as ecological psychology, which
>ph> explicitly reject it. People who have been trained to perform
>ph> little experiments to test (often rather silly) little ideas
>ph> [cannot] lay [exclusive] claim to the use of words like 'hypothesis'.
>ph>
>ph> And in any case, the whole practice of AI can be regarded as the
>ph> empirical testing of the hypothesis. Of course those who are working
>ph> under its aegis do not constantly question it, but take it as an
>ph> assumption and see how much science can be developed under it.
>ph> That is the way that science makes progress, in fact, as Kuhn has
>ph> argued convincingly. The world has plenty of serious people who reject
>ph> the PSH and are using other frameworks to develop and test theories of
>ph> mentality, and a large number of vocal and argumentative critics, so
>ph> there is no risk of its not being tested.
>ph>
>ph> Turning now to your second paragraph. You accuse Dyer of making
>ph> arguments which are 'in principle possible but in practice right
>ph> out of the window'. This, in a discussion which flows from a
>ph> hypothesis in which a human being memorises the code of a
>ph> program which can pass the Turing Test in Chinese, while preserving
>ph> his equanimity to the extent that he can simultaneously discuss
>ph> the code! If we are to reject unrealistic examples, then we can
>ph> all surely agree that the whole matter is a complete waste of
>ph> time, and just forget about it, starting now.
>ph>
>ph> Pat Hayes

PSH is certainly an empirical hypothesis if it is construed as a hypothesis about how "cognitive" engineers might successfully generate mind-like performance computationally (and people may differ in their judgments about how successful computation has been in doing that so far). But PSH is more like an untestable conjecture if is construed as the claim that the successful generators of that mind-like performance (if there are any) will have real minds (i.e., somebody will be at home in there), because normally the only way to know whether or not a system has a mind is to BE the system. Hence, for the very same reason that one can suppose that a stone (or any other body other than one's own) does or does not have a mind, as one pleases, without any hope of ever being any the wiser, the PSH is shielded from refutation by the impenetrability of the other-minds barrier.

Now Searle has figured out a clever way (I've dubbed it "Searle's Periscope") in which he could peek through the other-minds barrier and BE the other system, thus testing what would normally be an untestable conjecture. Searle's Periscope works ONLY for the special case of PSH (implementation-independent symbol manipulation): He has simply pointed out that if we (1) SUPPOSE (arguendo) that a physical symbol system alone could pass the Turing Test in Chinese, and from this we wish to (2) INFER that that physical symbol system would therefore be understanding Chinese (purely in virtue of implementing the TT-passing symbol system), THEN it is intuitively obvious that if (3) Searle himself implemented that same symbol system by memorizing all the symbols and rules and then performing the same symbol manipulations on the same inputs, then (4) he would NOT be understanding Chinese; therefore the inference to (2) (and hence the PSH) is false.

What makes this example unrealistic is much more the supposition (1) that a symbol system could pass the TT (there's certainly no such system in empirical sight yet!) rather than (3) that (if so, then) Searle could himself memorize and perform the same symbol manipulations. So maybe life is too short and memory too weak for a person to memorize and perform all those symbols and rules: So memorize and perform a few of them, and then a few more, and see if that kind of thing gives you a LITTLE understanding of Chinese! What is intuitively obvious is that there's nothing in the scenario of doing THAT kind of mindless thing till doomsday that would even faintly justify believing that that's the road to understanding.

No, the real sci-fi in this example comes from (1), not (3); and dwelling instead on the unrealistic features of (3) is motivated only by the yearning to re-establish the barrier that normally makes it impossible to block the conjecture that a system other than oneself has (or does not have, as the case may be) a mind. Mike Dyer tries to resurrect the barrier by supposing that Searle would simply develop multiple-personality syndrome if he memorized the symbols and rules (but why on earth would we want to believe THAT); you, Pat, try to resurrect the barrier by denying that Searle would really be a valid implementation of the same symbol system despite passing the same TT, using the same symbols and rules! And in response to "why not?" you reply only that his free will to choose whether or not to follow the rules is what disqualifies him. (Actually, I think it's his capacity to talk back when we project the PSH conjecture onto him that's the real problem; because that's something the poor, opaque first-order physical symbol system, slavishly following the very same rules and passing the same TT, is not free to do, any more than a stone is.)

Still others try to resurrect the other-minds barrier by invoking the fact that it is unrealistic to suppose that Searle could have the speed or the memory capacity to implement the whole symbol system (as if somewhere in the counterfactual realm of greater memory and greater speed there would occur a phase transition into the mental!).

To my mind, all these strained attempts to reject (4) at all costs are simply symptomatic of theory-saving at a mounting counterfactual price. I, like Tim Smithers, simply prefer taking the cheaper (and, I think, more realistic and down-to-earth) road of grounded robotics, abandoning pure computation, PSH, and the expanding ring of epicycles needed to keep them impenetrable to Searle's Periscope.

Stevan Harnad

Date: Thu, 23 Apr 92 00:34:42 PDT From: Dr Michael G Dyer Subject: physicality

Here are some responses and comments to whoever is willing to read them:

>
\>sh> But if the drowning is "virtual" (i.e., a computer-simulated
>
\>sh> person is "drowned" in computer-simulated water)
>
\>sh> there is no drowning at all going on, no matter how formally
>
\>sh> equivalent the symbols may be to real drowning.

I agree that there's no physical drowning, but what if we build an artifical neural network circuitry (with ion flow and/or action potential timings identical to those of some person's brain, etc.) and then give it the same inputs that a drowning person would receive? Who is to say that this artificial neural network won't have the subjective experience of drowning?

>
\>sh> ... as an intuition pump you can
>
\>sh> implement pain or proprioception by computation alone to the
>
\>sh> same degree that you can implement flying or heating by
>
\>sh> computation alone.)...
>
\>sh> I just don't think computation alone can either fly or think.

Here is where my intuition pumps diverges quite sharply. Is the physical act of something flying thru the air a computation? I think not (unless we imagine the entire universe as being a simulation God's computer -- then it is, but we'll never know :-). But does the EXPERIENCE of flying fall into a certain class of computations? No one really knows, but my bet is "yes". In that case, the actually physical act of flying is irrelevant. For a mind, what is important is the experience of flying.

I think that certain classes of computations actually have subjective inner experiences. At this point in time science simply has no way of even beginning to formulate a "theory" of what the subjective-point-of-view might be like for different types of computations, whether in VLSI, on tapes, optical or biochemical. Given that we can't tell, the safest strategy is to make judgements about the inner life of other entities based on their behavior.

ts> First, the actual practice of (symbol processing) AI research ts> makes it very difficult to talk about the Physical Symbol System ts> Hypothesis (PSSH) of Newell and Simon as being "a working ts> hypothesis". It is much more a widely accepted and unquestioned ts> dogma than it is a hypothesis. For it to be a hypothesis, in the ts> normal scientific sense (symbol processing) AI research would ts> need to be conducting a broad experimental investigation of its ts> validity (or otherwise). Very little, if any, research is either ts> presented as, or can be properly understood to be, a contribution ts> to such a research programme.

It is common for paradigm-level hypotheses to go unquestioned by those who are working within that paradigm (i.e. they accept the hypothesis and so don't spend time questioning or re-questioning it.)

In this context I think that Harnad and Searle play a very useful role in forcing some of us (more philosophically oriented) AI researchers to reexamine this hypothesis.

ts> ...does Mike Dyer think that they are going to be convinced by such ts> in principle possibly true but in practice right out of the window ts> aguments as he offers?

Mmmm.... so it's ok to have Searle do all symbol manipulations (that might require a level of granularity where each symbol represents a synapse or something lower) all in his head(!), but it's NOT ok for me to examine how one network (i.e. Searle's neurons) might be intertwined with another network (i.e. artificial VLSI circuitry isomorphic to the symbol relations and manipulations that make up a Chinese persona)??? My students and I happen to design connectionist-style networks of various sorts, to process language, make inferences, etc. and the issue of how one network gates and/or is composed with another etc. we think is rather relevant to understanding ultimately how minds might reside in brains.

Tough questions, however, are: "What's it's feel like to BE a particular sort of network?" "What's it feel like to BE a particular sort of (software) system? Harnad and Searle seem to assume that, no matter how complex, any kind of software system has no feelings. How do they know? Harnad claims we can simply ASK Searle to find out what it's like to understand English but he won't allow us to simply ASK the Chinese persona to find out what it's like to understand Chinese.

ts> I think all you guys should spend some time trying to build real ts> robots (not simulations!) that reliably do real things (even very ts> simple things) in the real world.

Yes, that's laudable, but we can't all be roboticists. However, I just saw one of those history of computer science shows and they had a nice demonstration of "virtual reality". VR is getting pretty good. You tilt the helmet and quite realistic images get updated, with proper perspective, etc. What if the robot received visual input from a VR world rather than the real world? (Oops! There goes those vision input transducers Stevan Harnad needs so badly! :-)

>
\>sh> By the way, minds and computers may both respond to syntax,
>
\>sh> but only minds respond to meaning. Computers are merely
>
\>sh> INTERPRETABLE as if they responded to meaning...

This is quite a claim! What evidence is there for such a claim? In contrast, each neuron appears to respond to its inputs (including its local chemical environment) without requiring any sort of thing called "meaning". The term "meaning", as far as I can tell, is simply used to refer to incredibly complex syntactic types of operations. If a robot (or person) is organized to behave in certain, very complex ways, then we tend to take (as Dennett says) an "intentional stance" toward it, but that doesn't mean there is anything other than syntax going on. (Biologists have also abandoned "life force" notions for the incredibly complex but syntactic operations of biochemistry.) The notion of "meaning" is useful for human-human folk interactions but the hypothesis of AI (and cognitive science in general) is that "meaningful" behavior is the result of a great many "mindless" (i.e. syntactic) operations (whether they are directly in circuitry or data structures in the memory of more general intepretation circuitry).

A simple example of meaning-from-syntax is the use of state space heuristic search (totally syntactic) to give an overall behavior that one might call "purposive" (e.g. a chess playing program "wanting" to checkmate its opponent).

The only "evidence" for meaning is probably from introspection. Of course, I can (and do) describe myself as having "meanings" because I can use that word to describe a certain class of complex behaviors and I happen to also exhibit those complex behaviors. But because I describe myself in this way does not require that I actually have some magical "meanings" that are something other than syntactic operations. Likewise, any really complex robot, capable of forming models of itself and of others, will take an intentional stance, both toward those other complex agents, AND toward itself -- i.e. attributing "meanings" to itself! So what? It's all ultimately just syntax. The trick is to figure out what class of marvelously complex syntactic operations brings about behaviors that deserve the folk psychological term of "meaning". (This reductionist approach in cognitive science is similar to that in the physical/natural sciences.)

>
\>sh> Searle simply points out to us that if he
>
\>sh> himself implemented the program (by memorizing the symbols
>
\>sh> and symbol manipulation rules) he would not understand Chinese,
>
\>sh> hence neither would any computer that implemented the same
>
\>sh> program.

Ask the LISP interpreter (that's executing code that creates some natural language understanding system S) if it "understands" anything and, of course, it doesn't. Ask S, however, and you will get an answer. We don't expect the LISP interpreter to "understand" what it's doing, so why should we EXPECT Searle to understand Chinese??? However, if we ask the Chinese persona what it's like to understand Chinese we will get an answer back (in Chinese).

For all my disagreements with Harnad, I think that he is concerned with an extremely interesting question, namely, what is the role that physicality plays in cognition? As we know, two "functionally identical" computations on two machines with different architectures are only similar in their computations at some level of abstraction. Below that level of abstraction, what the machines are doing physically may be very different. AI researchers believe that "consciousness" can (ultimately) reside on many different physical substrata as long as the computations are similar at some (as yet unspecified) level of abstraction and this level of abstraction can be modeled by symbol manipulation. The support for this view is that there seems to be no limit to the granularity of symbols and symbol manipulation (i.e. they can be made to correspond to the foldings of individual proteins if these are deemed essential in constructing the operations of a mind). Also, since we can only judge intentionality via behavior, pragmatically we never have to consider any level of abstraction below that level of computation that gives us behavior that appears intentional.

One final comment. There are two different uses of the term "grounding": 1. that representations should be rich enough to encode any perceptual information. 2. that physical transducers are required for intentionality.

I accept 1. but not 2. If a simulated robot could pass the TTT test within a virtual reality world, it would be grounded in that world but there would be no physical transducers. (I have never figured out why Harnad rejects out of hand the possibility of a "brain in a vat" whose I/O channels are wired up to a computer so that the brain thinks it’s seeing, standing, etc. Perhaps he rejects it because, if he doesn’t, then his whole requirement for physical transducers falls apart.)

Since Harnad plugs his own writings in this area, I will also:

Dyer, M. G. Intentionality and Computationalism: Minds, Machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence, Vol. 2, No. 4, 1990.

Dyer, M. G. Finding Lost Minds (Author’s reply to S. Harnad’s "Lost in the Hermeneutic Hall of Mirrors"). Journal of Experimental and Theoretical Artificial Intelligence, Vol. 2, No. 4, 1990.

Date: Thu, 23 Apr 92 20:21:46 EDT From: "Stevan Harnad"

ON MEANING: VIRTUAL AND REAL

I will preface my reply to Mike Dyer with a few points that should be kept in mind in my response:

(1) By "meaning," I mean subjective meaning, e.g., what it is like for a real English speaker (and not for a non-English-speaker or a stone) to hear and understand what spoken English is about. When I say that there is no real meaning in a symbol system, just symbols that are systematically interpretable as if they meant something, I mean that subjective meaning is absent: Either nobody is home at all (as in a stone) or certain particular symbols happen to have no subjective meaning for the system (as in Searle’s Chinese Room).

(2) "Grounding" is a robot’s capacity to interact with the real world of objects, events and states of affairs that its symbols are interpretable as being about. The semantic interpretations of the symbols and the robotic interactions with the objects must cohere TTT-indistinguishably with one another. This means that the symbol use is not constrained by syntax alone.

Grounding is not provably a necessary or a sufficient condition for subjective meaning (though it may in reality be necessary).

(3) The trouble with "brains in vats" is that they equivocate between (a) real de-afferented brains (with sensory surfaces removed, but all the rest of the neural hardware -- most of it analog -- intact) and (b) physical symbol systems (computers without any peripheral I/O devices). These two are radically different, and projecting assumptions about one onto the other leads nowhere. Brain-in-vat arguments usually further equivocate on the two senses of internal/external discussed earlier: inside/outside the "body" and inside/outside the "mind." Apart from the insistence on not conflating any of these things, I have no objections to brain-in-vat talk.

(4) As in (3) above, I insist on maintaining the distinction between real physical objects (like planes, furnaces, neurons and transducers) and their "virtual" counterparts (computer-simulated planes, etc.), be they ever so computationally equivalent to one another. It is the perpetual blurring of this boundary in particular that leaves me no choice but to keep repeating boringly to Mike Dyer that he seems to be hopelessly lost in a hermeneutic hall of mirrors he has created by overinterpreting systematically interpretable computations and then reading off the systematic interpretations themselves by way of evidence that the virtual world is as as real as the real one.

Michael G Dyer wrote:

md> what if we build an artificial neural network circuitry (with ion flow md> and/or action potential timings identical to those of some person’s md> brain, etc.) and then give it the same inputs that a drowning person md> would receive? Who is to say that this artificial neural network won’t md> have the subjective experience of drowning?

Is this a purely computational simulation of a neural network (i.e., a bunch of squiggles and squoggles that are interpretable as if they were ions, action potentials, etc.)? Or is a synthetic neural network with the same causal powers as the the organic neural network (i.e., the capacity to transduce all the same real physical input the neural network gets)? If it’s the former, then it’s really all just squiggles and squoggles, no matter how you can systematically interpret it. If it’s the latter then it’s a real artificial neuronal circuit and can in principle have real subjective experiences (but then it’s irrelevant to this discussion of pure computation and whether or not pure computation can have subjective experiences).

md> Is the physical act of something flying thru the air a computation? I md> think not... But does the EXPERIENCE of flying fall into a certain md> class of computations? No one really knows, but my bet is "yes". I md> think that certain classes of computations actually have subjective md> inner experiences.

The trouble with experiences is that all but your own are out of sight, so you are free to interpret any external object or process as if it had experiences. Trouble is, there’s a right and wrong of the matter (even though the other-minds barrier normally prevents us from knowing what it is). There were reasons, for a while, for entertaining the hypothesis that experiences might be implemented computations. A lot of this discussion is about reasons why that hypothesis has to be reconsidered and discarded.

md> [why is it] ok to have Searle do all symbol manipulations (that might md> require a level of granularity where each symbol represents a synapse md> or something lower) all in his head(!), but... NOT ok for me to md> examine how one network (i.e. Searle’s neurons) might be intertwined md> with another network (i.e. artificial VLSI circuitry isomorphic to the md> symbol relations and manipulations that make up a Chinese persona)?

Until further notice, real neurons have nothing to do with this. What Searle and the TT-passing computer he is duplicating are doing is implementation-independent. We don't know what the real brain does; let us not presuppose anything. Nor do I know what you are imagining intertwining: virtual neurons and what? It's all just squiggles and squoggles!

[Here's a prophylactic against hermeneutics: "When in certainty, de-interpret all symbols and see what's left over."]

md> Harnad and Searle seem to assume that, no matter how complex, any kind md> of software system has no feelings. How do they know? Harnad claims we md> can simply ASK Searle to find out what it's like to understand English md> but he won't allow us to simply ASK the Chinese persona to find out md> what it's like to understand Chinese.

Because of the other-minds problem we cannot KNOW that anyone else but ourselves has feelings (or no feelings, as the case be). I am prepared to believe other people do, that animals do, and that various synthetic systems might too, particularly TTT-scale robots. I'm even prepared to believe a computer might (particularly since I can't KNOW that even a stone does not). There is only one thing I am not prepared to believe, and that is that a computer has feelings PURELY IN VIRTUE OF RUNNING THE RIGHT PROGRAM (i.e., the physical symbol system hypothesis). But, unfortunately, that's precisely what's at issue here.

You fault me for believing Searle (and his quite reasonable explanation of what is going on -- meaningless symbol manipulation) rather than the Chinese squiggles and squoggles. But you are prepared to believe that Searle has gotten multiple personality merely as a consequence of having memorized and performed a bunch of symbol manipulations, just because of what the symbols are interpretable as meaning.

Finally (although I don't want to push the premise that such a TT-passing computer program is even possible too hard, because we've accepted it for the sake of argument), you don't seem too troubled by the fact that the Chinese "persona" couldn't even tell you what Searle was wearing at the moment. Any self-respecting multiple personality could manage that. Doesn't this suggest that there might be a bit more to real-world grounding and the TTT than is apparent from the "just ask the simulation" perspective?

md> What if the robot received visual input from a VR world rather than the md> real world? ... There go those visual input transducers Stevan Harnad md> needs so badly!)

Real robots have real sensory surfaces. I have no objection to those real sensory surfaces being physically stimulated by stimulation generated by a simulated world, itself generated by a computer. (My robot would then be like a kid sitting in a driving simulator.) That would show nothing one way or the other. But please don't talk about de-afferenting my robot and reducing him to a "brain-in-vat" and then piping the computer-generated input straight to THAT, because, as I said before, neither of us knows what THAT would be, To assume otherwise (e.g., that it would be a computer) is simply to beg the question!

md> each neuron appears to respond to its inputs (including its local md> chemical environment) without requiring any sort of thing called md> "meaning". The term "meaning", as far as I can tell, is simply used to md> refer to incredibly complex syntactic types of operations. [If] a robot md> (or person) is organized to behave in certain, very complex ways, then md> we tend to take (as

Dennett says) an "intentional stance" toward it, md> but that doesn't mean there is anything other than syntax going on.

Being interpretable (by an outsider) as having subjective meaning, no matter how practical or useful, is still not the same as (and certainly no guarantor of) actually having subjective meaning. Subjective meaning does NOT simply refer to "incredibly complex syntactic types of operations"; and, as usual, neurons have nothing to do with this (nor are their activities "syntactic"). And where subjective meaning is going on there is definitely more than (interpretable) syntax going on.

md> The only "evidence" for meaning is probably from introspection. Of md> course, I can (and do) describe myself as having "meanings" because I md> can use that word to describe a certain class of complex behaviors and md> I happen to also exhibit those complex behaviors. But because I md> describe myself in this way does not require that I actually have some md> magical "meanings" that are something other than syntactic operations.

You don't really understand English and fail to understand Chinese because you "describe [yourself] as having 'meanings'" but because there's real subjective understanding of English going on in your head, along with real subjective experience of red, pain, etc. Besides really experiencing all that, you're also describable as experiencing it; but some systems are describable as experiencing it WITHOUT really experiencing it, and that's the point here! Explanatory convenience and "stances" -- by outsiders or by yourself -- have nothing whatsoever to do with it. There's nothing "magic" about it either; just something real!

md> Ask the LISP interpreter (that's executing code that creates some md> natural language understanding system S) if it "understands" anything md> and, of course, it doesn't. Ask S, however, and you will get an answer. md> We don't expect the LISP interpreter to "understand" what it's doing, md> so why should we EXPECT Searle to understand Chinese??? However, if we md> ask the Chinese persona what it's like to understand Chinese we will md> get an answer back (in Chinese).

Taking S's testimony about what it's like to understand Chinese as evidence against the claim that there is no real subjective understanding going on in there is like taking the fact that it "burns" (simulated) marshmallows as evidence against the claim that a (simulated) fire is not really hot. This is precisely the type of hermeneutic credulity that is on trial here. One can't expect to gain much credence from simply citing the credulity in its own defense (except from someone else who is caught up in the same hermeneutic circle).

md> If a simulated robot could pass the TTT test within a virtual reality md> world, it would be grounded in that world but there would be no md> physical transducers. I have never figured out why Harnad rejects out md> of hand the possibility of a "brain in a vat" whose I/O channels are md> wired up to a computer so that the brain thinks it's seeing, standing, md> etc.

md> Michael G Dyer

Virtually grounded, not really grounded, because of course that's only a virtual TTT, not a real one. But the whole point of the TT/TTT distinction was to distinguish the merely virtual from the real!

Stevan Harnad

From: Pat Hayes Date: Wed, 22 Apr 92 14:25:29 MDT

>
>sh> the structures and processes
>
>sh> underlying our capacity to categorize inputs (beginning with sensory
>
>sh> projections).... will turn out to be
>
>sh> largely nonsymbolic, but perhaps symbols can be grounded in the
>
>sh> capacity those nonsymbolic structures and processes give us to pick out
>
>sh> the objects they are about.

If we include (as we should) linguistic input, it seems clear that the structures and processes will be largely symbolic. I think that vision and other perceptual modes involve symbols from an early stage, but I agree that's just one intuition against another.

I think there is something important (though vague) here:

>
>ph> Here's an example adapted from one of Brian [Smith's]. Take a set
>
>ph> of rules which encode (a formal system for) arithmetic, together with
>
>ph> a formal predicate 'lengthof', and the rules
>
>ph>
>
>ph> lengthof('0') -> 1
>
>ph> lengthof(n<>x) -> lengthof(n) + lengthof(x)
>
>ph>
>
>ph> Now, these rules make 'lengthof(n)' evaluate to (a numeral which means)
>
>ph> the number of digits in the formal representation of n: ie, the length
>
>ph> of that numeral in digits. Notice this is the ACTUAL length of that
>
>ph> piece of syntax. Now, is this 'formal'? It is according to your
>
>ph> definition, and perhaps you are happy with that, but it has some marks

>
>ph> which successfully refer to physical properties of part of the world. >
>
>sh> But note that in your example above, even though the computation yields
>
>sh> a symbol that is interpretable as the number of symbols in the string,
>
>sh> this is in principle no different from a computation that yields a
>
>sh> symbol that is interpretable as the number of planets in the solar
>
>sh> system. It is just a systematic correspondence (and hence interpretable
>
>sh> as such)

No, you have missed the point of the example. The difference is that in this example, the sytematicity is between the syntax of one numeral and the actual (physical?) length of another. This is not the same kind of connection as that between some symbols and a piece of the world that they can be interpreted as referring to. It requires no external interpreter to make it secure, the system itself guarantees that this interpretation will be correct. It is a point that Descartes might have made: I don't need to be connected to an external world in any way in order to be able to really count.

>
>sh> ... But "interpretable as meaning X" (as in the case of a book,
>
>sh> interpretable by a thinking mind) is not the same as "meaning X" (as in
>
>sh> the case of thoughts, in a mind). Failing to distinguish the two seems
>
>sh> to be another instance of conflating physical inner/outer and mental
>
>sh> inner/outer, as discussed earlier.

I am distinguishing them, and claiming to have a case of the latter. Now of course if you insist a priori that meaning can only take place in a mind, and a system like this isn't one, then you have the day; but that seems to beg the question.
>
>
>ph> My intuition tells me clearly that when I debug a piece of code by
>
>ph> pretending to be an interpreter and running through it "doing" what it
>
>ph> "tells" me to do, that the program is not being run, and certainly not
>
>ph> run on, or by, me. So we are left with your intuition vs. my intuition,
>
>ph> and they apparently disagree.
>

>
>sh> But isn't the real question whether there is any relevant difference
>
>sh> between what you think is a "real" implementation by a machine and what
>
>sh> you think is a "pseudo-implementation" by a person? Certainly the
>
>sh> computer is not stepping through the states consciously and
>
>sh> deliberately, as you are. But is there anything else that's different?
>
>sh> If we speak only of the "motions gone through" and their I/O conditions
>
>sh> in the two cases, they are exactly the same. In the case of the
>
>sh> machine, the motions are mechanical; no choice is involved. In the case
>
>sh> of the person, their elective. But so what?

Well, that is a very good question. That is exactly what computer science is all about. What is different in having a machine that can run algorithms from just being able to run algorithms? I take it as obvious that something important is, and that answering that question is, pace Brian Smith's recent message, essentially an empirical matter. We are discovering so what.

>
>sh> Even apart from the vexed
>
>sh> questions associated with free will and causality, what is there about
>
>sh> taking IDENTICAL motions under identical I/O conditions and making
>
>sh> their causal basis mindless and mechanical that could possibly effect a
>
>sh> transition INTO the mental (rather than OUT of it, which is the much
>
>sh> more obvious feature of the transition from the human implementation to
>
>sh> the machine one)?

I agree it seems almost paradoxical. But as I emphasised, the key is that these ARENT identical sequences of states. Thats what computers do. They put algorithms into the physical world, give them a life of their own, enable them to become real in some important sense. Its a hard sense to get exactly clear, but it seems very real. The difficulty is illustrated well by the awful trouble software is giving to legal concepts, for example. Since they are textual and can be copied, and do nothing until 'performed', they seem like things to be copyrighted. But in many ways they are more like pieces of machinery suitable for patenting. They are both, and neither: they are something new.

>
>sh> It’s always useful, in this sort of hermeneutic puzzle, to de-interpret
>
>sh> and reduce things to gibberish as much as possible

Ah, maybe that is a bad heuristic sometimes. Clearly if you insist that this can always be done to computer insides but not always to human insides, then you are never going to see meaning in a machine.

>
>sh> Suppose the computer was doing all the requisite
>
>sh> summation in binary, and you were too,
>
>sh> and all it did, and all you did, was compare zero’s and one’s and erase
>
>sh> and carry, just like a Turing Machine. Is it still so obvious that
>
>sh> you’re not doing everything the computer is doing? If anything, the
>
>sh> computer is doing less than you rather than more (because it has no
>
>sh> choice in the matter). Why should I interpret less as more?

The computer is doing less than me, but thats my point: the PROGRAM is more responsible for what is happening. The computer is essentially BECOMING the program, one might almost say, giving its symbolic patterns momentary flesh so that they act in the world. And thats what a human reader of the code is not doing (unless hypnotised or somehow in its grip in some unreal way).

>
>sh> By the way, minds and computers may both respond to syntax, but only
>
>sh> minds respond to meaning. Computers are merely INTERPRETABLE as if they
>
>sh> responded to meaning...

Nah nah, question begging again!

>
>sh> And I bet a lot of the essential features of pain and proprioception
>
>sh> will be in the analog properties of the hardware that implements it,
>
>sh> which will be more like exotic chemistry.

OK, last word is yours. Who is taking the bets?

Pat Hayes

After a brief lull (mainly because I was out of town and fell behind with the postings) the "What is Computation" discussion proceeds apace... -- SH

Date: Wed, 29 Apr 1992 22:20:04 -0400 From: mcdermott-drew@CS.YALE.EDU (Drew McDermott)

I will respond to both Stevan Harnad and Pat Hayes; I wrote:

>dm> Let's distinguish between a computer's states' being
>dm> "microinterpretable" and "macrointerpretable." The former case is what
>dm> you assume: that if we consider the machine to be a rewrite system, the
>dm> rewrite rules map one coherently interpretable state into another. Put
>dm> another way, the rewrite rules specify a change in belief states of the
>dm> system. By contrast, the states of a macrointerpretable system "sort of
>dm> line up" with the world in places, but not consistently enough to
>dm> generate anything like a Tarskian interpretation. What I think you've
>dm> overlooked is that almost all computational processes are at best
>dm> macrointerpretable.

Pat Hayes replied:

>ph> Drew, clearly you have an antisemantic axe to grind, but its not
>
>sh> very sharp.

I do have axes to grind, but this isn't one of them. I do not dispute that computers do normally succeed in referring to things and states to exactly the same degree that we do. But the question at issue is whether this fact is part of the *definition* of "computer." I'm pretty sure that Pat and I agree here: that computers are defined as physical instantiations of formal automata (I won't repeat David Chalmers's excellent statement of the position), and they happen to make excellent semantic engines when connected up to things their states can come to refer to.

Now back to Stevan:

You raise four semi-independent issues:

>
>sh> (1) Does EVERY computer implementing a program have SOME states that are
>
>sh> interpretable as referring to objects, events and states of affairs, the
>
>sh> way natural language sentences are?

>
>sh> (2) Are ALL states in EVERY computer implementing a program interpretable
>

>sh> as referring... (etc.)?

>
>sh> (3) What is the relation of such language-like referential
>
>sh> interpretability and OTHER forms of interpretability of states of a
>
>sh> computer implementing a program?

>
>sh> (4) What is the relation of (1) - (3) to the software hierarchy, from
>
>sh> hardware, to machine-level language, to higher-level compiled
>
>sh> languages, to their English interpretations?

>
>sh> My answer would be that not all states of a computer implementing a
>
>sh> program need be interpretable, and not all the interpretable states
>
>sh> need be language-like and about things in the world (they could be
>
>sh> interpretable as performing calculations on numbers, etc.), but ENOUGH
>
>sh> of the states need to be interpretable SOMEHOW, otherwise the computer
>
>sh> is just performing gibberish (and that's usually not what we use
>
>sh> computers to do, nor do we describe them as such), and THAT's the
>
>sh> interpretability that's at issue here.

But it isn't! We're talking about whether semantic interpretability is part of the *definition* of computer. For that to be the case, everything the computer does must be semantically interpretable. Does it cease to be a computer during the interludes when its behavior is not interpretable?

I assumed that your original claim was that a computer had to correspond to an interpreted formal system (where, in the usual case, the users supply the interpretation). But that's not what you meant at all. An interpreted formal system includes a mapping from states of the system to states of the world. Furthermore, there is a presumption that the state-transition function for the formal system preserves the meaning relation; if the state of affairs denoted by system state S1 holds, then the state of affairs denoted by the following state also holds. But now it's clear that neither you nor Pat is proposing anything of this sort. Instead, you seem to agree with me that a computer is a physical embodiment of a formal automaton, plus a kind of loose, pragmatic, fault-prone correspondence between its states and various world states. Given this agreement, let's simplify. Clearly, the semantic interpretation is no part of the definition of computer. We can identify computers without knowing what interpretation their users place on them.

I have lots more examples. Today I saw a demo of a computer generating the Mandelbrot set. (It was the DEC alpha chip; definitely the Mandelbrot engine of choice.) Unquestionably a computer; what did its states denoteg It seems clear at first: The color of each pixel denoted a speed of convergence of a certain numerical process. But that's just the platonic ideal. But platonic referents are very unsatisfactory for our purposes, on two counts. (1) If we count platonic referents, then *any* formal system has a trivial set of referents. (2) The viewer of the screen was not interested in this set of referents, but in the esthetic value of the display. Hence the real universe of the users was the universe of beauty and truth. Vague, of course, but *computers' semantic relations are normally vague.*

>dm> More examples: What do the states of a video game refer to? The Mario
>dm> brothers? Real asteroids?

>
>sh> They are interpretable as pertaining (not referring, because there's no
>
>sh> need for them to be linguistic) to (indeed, they are hard-wireable to)
>
>sh> the players and moves in the Mario Brothers game, just as in chess. And
>
>sh> the graphics control component is interpretable as pertaining to (and
>
>sh> hard-wireable to the bit-mapped images of) the icons figuring in the
>
>sh> game. A far cry from uninterpretable squiggles and squoggles.

The "players and moves" mostly don't exist, of course, since they include entities like King Koopa and Princess Toadstool. The child playing the game thinks (sort of) that the pictures on the screen refer to a certain universe. Or maybe they *constitute* a universe. It's hard to be precise, but I hope by now vagueness doesn't bother you. Of course, the engineer that wrote the game knows what's *really* going on. The input signals refer to presses of control buttons by the game player. Output signals refer to shapes on a screen. But it would be absurd to say that the game player's version of the semantics is only an illusion, and the real purpose of the system is to map buttons pushes onto screen alterations. Shall we say, then, that there are *two* computers here --- one formal system, but two competing semantic interpretationsg I'd rather say that there is one computer, and as many interpretations as are convenient to posit --- including possibly zero. [Also, the engineer's interpretation is almost trivial, because all it refers to are the program's own inputs and outputs; almost, but not quite, because normally the inputs are real pressures on buttons and the outputs are real photons emanating from a screen.]

>dm> Take almost any example, a chess program, for instance. Suppose that
>dm> the machine is evaluating a board position after a hypothetical series
>dm> of moves. Suppose the evaluation function is a sum of terms. What does
>dm> each term denote? It is not necessary to be able to say. One might, for
>dm> instance, notice that a certain term is correlated with center control,
>dm> and claim that it denotes "the degree of center control," but what does
>dm> this claim amount to? In many games, the correlation will not hold, and
>dm> the computer may as a consequence make a bad move. But the evaluation
>dm> function is "good" if most of the time the machine makes "good moves."

>
>sh> I’m not sure what an evaluation function is,

[It’s the program that computes a quick guess of how good a board position is without any further lookahead.]

>
>sh> but again, I am not saying
>
>sh> every state must be interpretable. Even in natural language there are
>
>sh> content words (like "king" and "bishop") that have referential
>
>sh> interpretations and function words ("to" and "and") that have at best
>
>sh> only syntactic or functional interpretations. But some of the internal
>
>sh> states of a chess-plying program surely have to be interpretable as
>
>sh> referring to or at least pertaining to chess-pieces and chess-moves, and
>
>sh> those are the ones at issue here.

But if only "some" of the states have to be interpretable, then is the system only a computer some of the timeg Or to some degree?

>dm> The chess program keeps a tree of board positions. At each node of this
>dm> tree, it has a list of moves it is considering, and the positions that
>dm> would result. What does this list denote? The set of moves "worth
>dm> considering"? Not really; it’s only guessing that these moves are worth
>dm> considering. We could say that it’s the set the machine "is
>dm> considering," but this interpretation is trivial.

>
>sh> And although I might make that interpretation for convenience in
>
>sh> describing or debugging the program (just as I might make the
>
>sh> celebrated interpretation that first got Dan Dennett into his
>
>sh> "intentional stance," namely, that "the computer thinks it should get
>
>sh> it’s queen out early"), I would never dream of taking such
>
>sh> interpretations literally: Such high level mentalistic interpretations
>
>sh> are simply the top of the as-if hierarchy, a hierarchy in which
>
>sh> intrinsically meaningless squiggles and squoggles can be so interpreted

>
>sh> that (1) they are able to bear the systematic weight of the
>
>sh> interpretation (as if they "meant" this, "considered/believed/thought"
>
>sh> that, etc.), and (2) the interpretations can be used in (and even sometimes
>
>sh> hard-wired to) the real world (as in interpreting the squiggles and
>
>sh> squoggles as pertaining to chess-men and chess-moves).

You're forgetting which side of the argument you're on. *I'm* arguing that such interpretations are epiphenomenal. *You're* arguing that the interpretation is the scaffolding supporting the computerhood of the system. Or perhaps I should picture a trapeze; if the system spends too much time between interpretable states, it falls from computational grace.

>dm> We can always impose a trivial interpretation on the states of the
>dm> computer. We can say that every register denotes a number, for
>dm> instance, and that every time it adds two registers the result denotes
>dm> the sum. The problem with this idea is that it doesn't distinguish the
>dm> interpreted computers from the uninterpreted formal systems, because I
>dm> can always find such a Platonic universe for the states of any formal
>dm> system to "refer" to. (Using techniques similar to those used in
>dm> proving predicate calculus complete.)

>
>sh> I'm not sure what you mean, but I would say that whether they are
>
>sh> scratches on a paper or dynamic states in a machine, formal symbol
>
>sh> systems are just meaningless squiggles and squoggles unless you project
>
>sh> an interpretation (e.g., numbers and addition) onto them.

At this point you seem to have crossed over and joined my side completely. You are admitting that there can be machines that embody formal symbol systems whose states are just meaningless squiggles and squoggles.

>
>sh> The fact that
>
>sh> they will bear the systematic weight of that projection is remarkable
>
>sh> and useful (it's why we're interested in formal symbol systems at all),
>
>sh> but certainly not evidence that the interpretation is intrinsic to the
>
>sh> symbol system;

Yes! Right!

>
\>sh> it is only evidence of the fact that the system is
>
\>sh> indeed a nontrivial symbol system (in virtue of the fact that it is
>
\>sh> systematically interpretable). Nor (as is being discussed in other
>
\>sh> iterations of this discussion) are coherent, systematic "nonstandard"
>
\>sh> alternative interpretations of formal symbol systems that easy to come
>
\>sh> by.

You're going to feel terrible when you realize you've agreed with me!

\>dm> If no other argument convinces you, this one should: Nothing prevents
\>dm> a computer from having inconsistent beliefs. We can build an expert
\>dm> system that has two rules that either (a) cannot be interpreted as
\>dm> about medical matters at all; or (b) contradict each other. The system,
\>dm> let us say, happens never to use the two rules on the same case, so
\>dm> that on any occasion its advice reflects a coherent point of view.
\>dm> (Sometimes it sounds like a homeopath, we might say, and sometimes like
\>dm> an allopath.) We would like to say that overall the computer's
\>dm> inferences and pronouncements are "about" medicine. But there is no way
\>dm> to give a coherent overall medical interpretation to its computational
\>dm> states.

>
\>sh> I can't follow this: The fact that a formal system is inconsistent, or
>
\>sh> can potentially generate inconsistent performance, does not mean it is
>
\>sh> not coherently interpretable: it is interpretable as being
>
\>sh> inconsistent, but as yielding mostly correct performance nevertheless.
>
\>sh> [In other words, "coherently interpretable" does not mean
>
\>sh> "interpretable as coherent" (if "coherent" presupposes "consistent").]

It matters in the traditional framework I was assuming you endorsed. I see that you don't. Pat does, however:

\>ph> Your inconsistent-beliefs point misses an important issue. If that
\>ph> expert system has some way of ensuring that these contradictory rules
\>ph> never meet, then it has a consistent interpretation, trivially: we can
\>ph> regard the mechanism which keeps them apart as being an encoding of a

>ph> syntactic difference in its rule-base which restores consistency.
>ph> Maybe one set of rules is essentially written with predicates with an
>ph> "allo-" prefix and the others with a "homeo-". You might protest that
>ph> this is cheating, but I would claim not: in fact, we need a catalog of
>ph> such techniques for mending consistency in sets of beliefs, since
>ph> people seem to have them and use them to 'repair' their beliefs
>ph> constantly, and making distinctions like this is one of them (as in,
>ph> "Oh, I see, must be a different kind of doctor"). If on the other hand
>ph> the system has no internal representation of the distinction, even
>ph> implicit, but just happens to never bring the contradiction together,
>ph> then it is in deep trouble ....

I'm with Stevan on this one. The rule-separation mechanism may in some sense restore consistency, but it's hard to explain how it does this *semantically.* (The syntactic mechanism must somehow affect the meanings of the rules, or affect the sense in which the system "believes" its rules.) Fortunately, we are not called on to provide a systematic semantics.

>dm> I suspect Searle would welcome this view, up to a point. It lends
>dm> weight to his claim that semantics are in the eye of the beholder.
>dm> ... However, the point
>dm> at issue right now is whether semantic interpretability is part of the
>dm> definition of "computer." I argue that it is not; a computer is what
>dm> it is regardless of how it is interpreted. I buttress that
>dm> observation by pointing out just how unsystematic most interpretations
>dm> of a computer's states are. However, if I can win the argument about
>dm> whether computers are objectively given, and uninterpreted, then I
>dm> can go on to argue that unsystematic interpretations of their states
>dm> can be objectively given as well.

>
>sh> If you agree with Searle that computers can't be distinguished from
>
>sh> non-computers on the basis of interpretability, then I have to ask you
>
>sh> what (if anything) you DO think distinguishes computers from
>
>sh> non-computers?

I refer you to Chalmers. A brief summary: A system is a computer if its physical states can be partitioned into classes that obey a transition relation.

Drew McDermott

Date: Thu, 7 May 92 19:01:34 EDT From: "Stevan Harnad"

ON IMPLEMENTING ALGORITHMS MINDLESSLY

Pat Hayes wrote:

>ph> If we include (as we should) linguistic input, it seems clear that
>ph> structures and processes [underlying our capacity to categorize] will
>ph> be largely symbolic... vision and other perceptual modes involve
>ph> symbols from an early stage...

The only problem with "including" (as you put it) linguistic input is that, without grounding, "linguistic input" is just meaningless squiggles and squoggles. To suppose it is anything more is to beg the main question at issue here.

To categorize is to sort the objects in the world, beginning with their sensory projections. It is true that we can sort names and descriptions too, but unless these are first grounded in the capacity to sort and name the objects they refer to, based on their sensory projections, "names and descriptions" are just symbolic gibberish that happens to have the remarkable syntactic property of being systematically translatable into a code that we are able to understand. But that's all MEDIATED meaning, it is not autonomously grounded. And a viable candidate for what's going on in our heads has to be autonomously grounded; it can't just be parasitic on our interpretations.

Another thing you might have meant was that symbols play a role even in sensory categorization. That may be true too, but then they better in turn be grounded symbols, otherwise they are hanging from a (Platonic?) skyhook.

>ph> No, you have missed the point of the [internal length] example.
>ph> in this example, the systematicity is between the
>ph> syntax of one numeral and the actual (physical?) length of another.
>ph> This is not the same kind of connection as that between some symbols
>ph> and a piece of the world that they can be interpreted as referring to.
>ph> It requires no external interpreter to make it secure, the system
>ph> itself guarantees that this interpretation will be correct. It is a
>ph> point that Descartes might have made: I don't need to be connected to
>ph> an external world in any way in order to be able to really count.

MENTAL counting is moot until its true underlying mechanism is known; you are simply ASSUMING that it's just symbol manipulation.

But your point about the correspondence between the internal numerical symbol for the length of an internal sequence can be made without referring to the mental. There is certainly a correspondence there, and the interpretation is certainly guaranteed by causality, but only in a slightly more interesting sense than the interpretation that every object can be taken to be saying of itself "Look, here I am!" That too is a guaranteed relation. I might even grant that it's "grounded," but only in the trivial sense that an arbitrary toy robot is grounded. Symbols that aspire to be the language of thought cannot just have a few fixed connections to the world. The systematicity that is needed has to have at least the full TT power of natural language -- and to be grounded it needs TTT-scale robotic capacity.

Arithmetic is an artificial language. As such, it is an autonomous formal "module," but it also happens to be a subset of English. Moreover, grounded mental arithmetic (i.e., what we MEAN by numbers, addition, etc.) is not the same as ungrounded formal arithmetic (symbols that are systematically interpretable as numbers).

That having been said, I will repeat what I said earlier, that there may nevertheless be something to learn from grounded toy systems such as the numerical one you describe. There may be something of substance in such dedicated systems that will scale up to the TTT. It's just not yet obvious what that something is. My guess is it will reside in the way the analog properties of the symbols and what they stand for (in this case, the physical magnitude of some quantity) constrain activity at the syntactic level (where the "shape" of the symbols is normally arbitrary and hence irrelevant).

>ph> What is different in having a machine that can run
>ph> algorithms from just being able to run algorithms? I take it as
>ph> obvious that something important is...

I think you're missing my point. The important thing is that the algorithm be implemented mindlessly, not that it be implemented mechanically (they amount to the same thing, for all practical purposes). I could in principle teach a (cooperative) two-year old who could not read or write to do rote, mechanical addition and multiplication. I simply have him memorize the finite set of meaningless symbols (0 - 9) and the small set of rules (if you see "1" above "3" and are told to "add" give "4", etc.). I would then have a little human calculator, implementing an algorithm, who didn't understand a thing about numbers, just as Searle doesn't understand a word of Chinese.

Now let me tell you what WOULD be cheating: If any of what I had the child do was anything but SYNTACTIC, i.e., if it was anything other than the manipulation of symbols on the basis of rules that operate only on their (arbitrary) shapes: It would be cheating if the child (mirabile dictu) happened to know what "odd" and "even" meant, and some of the calculations drew on that knowledge instead of just on the mechanical algorithm I had taught him. But as long it's just mechanical syntax, performed mindlessly, it makes no difference whatsoever whether it is performed by a machine or stepped through (mechanically) by a person.

Now if you want to appreciate the real grip of the hermeneutical circle, note how much easier it is to believe that an autonomous black box is "really" understanding numbers if it is a machine implementing an algorithm mechanically rather than an illiterate, non-numerate child, who is just playing a symbolic game at my behest. THAT's why you want to disqualify the latter as a "real" implementation, despite the fact that the same syntactic algorithm is being implemented in both cases, without any relevant, nonarbitrary differences whatsoever.

>ph> Clearly if you insist that [reducing to gibberish]
>ph> can always be done to computer insides but not always to human
>ph> insides, then you are never going to see meaning in a machine.

I am sure that whatever is REALLY going on in the head can also be deinterpreted, but you mustn't put the cart before the horse: You cannot stipulate that, well then, all that's really going on in the head is just symbol manipulation, for that is the hypothesis on trial here!

{Actually, there are two semi-independent hypotheses on trial: (1) Is anything NOT just a computer doing computation? and, (2) Are minds just computers doing computation? We agree, I think, that some things are NOT computers doing computation, but you don't think the mind is one of those noncomputational things whereas I do.]

I had recommended the exercise of deinterpreting the symbols so as to short circuit the persuasive influence of those properties that are merely byproducts of the interpretability of the symbols, to see whether there's anything else left over. In a grounded TTT-scale robot there certainly would be something left over, namely, the robotic capacity to discriminate, categorize and manipulate the objects, events and states of affairs that the symbols were about. Those would be there even if the symbols were just gibberish to us. Hence they would be grounding the interpretations independently of our mentalistic projections.

Stevan Harnad

Date: Thu, 7 May 92 19:23:41 EDT From: "Stevan Harnad"

To all contributors to the "What is Computation?" Symposium:

Jim Fetzer, Editor of the (paper) journal MINDS AND MACHINES has expressed interest in publishing the Symposium (see below) as a special issue of his journal. He has already published one such paper version of a "Skywriting" Symposium similar to this one, which will appear shortly as:

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on the Virtual Mind. Minds and Machines (in press)

That Symposium was smaller, with fewer participants, but I hope that the participants in this larger one will want to do this too. I will generate formatted hard copy, and the participants can polish up their prose and add references, but what we must avoid is pre-emptive re-writing that makes one another's contibutions retroactively obsolete. We also cannot coordinate diverging iterations of rewriting. We should instead preserve as much as possible the interactive "Skywriting" flavor of the real-time exchange, as we did in the other Symposium.

Please let me know which of you are (and are NOT) interested in publication. In the meanwhile, we can continue for a few more iterations before involking cloture. Perhaps the prospct of publication will change the style of interaction from this point on, perhaps not...

Several backed up postings are still waiting in the wings.

Bets wishes,

Stevan,

From: jfetzer@ub.d.umn.edu (james fetzer) Date: Wed, 22 Apr 92 18:20:34 CDT

Stevan,

This stuff is so interesting that I might devote a whole issue to it. How would you like to guest edit a special issue on the topic, "What is Computation?", for MINDS AND MACHINES? Beginning in 1993, we will be going to 125 pages per issue, and each page runs about 600 words. So that represents the maximum length of material I can use. If you like the idea, I have no deadline in mind, but I do believe that it may require something like position papers from the principal contributors in addition to exchanges. I am not envisioning what is nothing more than one continuous exchange, but I will be open-minded about any suggestions you may have about how we proceed.

Let me know if you like this suggestion. Jim

From: jfetzer@ub.d.umn.edu (james fetzer) Date: Thu, 30 Apr 92 16:09:12 CDT

On the proposed new skywriting project tentatively entitled, "What is Computation?", let's take things one step at a time. I like the project but I know we need to agree on a few ground rules.

(1) Let's start with a tentative length of 50 pages at 600 words per (30,000) and see how that plays. If you should need more, then we can work that out, but I would like to try 30,000 first.

(2) The authors must make appropriate reference to works that have been previously discussed or are otherwise relevant to their views. (Skywriting seems to invite unattributed use of ideas, etc., which both of us need to discourage.)

(3) The version that is submitted will be subject to review in accordance with the journal's standing policies. Such review may lead to revisions of certain parts of the exchange, but every effort will be made to adhere as closely as possible to the spirit of the original.

(4) When the final version is submitted to the publisher for typesetting, only typographical corrections will be allowed, lest we bog down in changes that generate other changes, over and over, due to the large number of contributors, etc.

(5) You will (once again) assume responsibility for preparing the manuscript for submission and will execute the permission to publish form on behalf of all of the contributors and will be responsible for overall proofing of the typeset manuscript, coordinating with the others as necessary.

If this sounds agreeable with you, then by all means, let us proceed. Keep me posted as things develop, but I would recommend that the number of contributors be kept to a managably small number, whatever that is.

Jim Fetzer

Editor MINDS AND MACHINES

Date: Fri, 8 May 92 13:10:51 EDT From: "Stevan Harnad"

Date: Thu, 7 May 92 19:18:06 EST From: David Chalmers To: harnad@Princeton.EDU Subject: Re: Publishing the "What is Computation" Symposium

Sounds fine, an interesting idea. I'll probably make one more contribution within the next week or so, addressing various points that have come up.

Cheers, Dave.

Date: Fri, 8 May 92 12:46:36 -0400 From: "John C. Haugeland" To: harnad@Princeton.EDU, jfetzer@ub.d.umn.edu Subject: Special issue of Minds and Machines on "What is Computation"

Dear Jim and Steve:

I have been following the discussion on "What is computation" with (predictable) interest, but I have not yet participated because of a backlog of prior commitments. These commitments will keep me preoccupied, alas, for the next six weeks as well -- mostly travelling. I have been intending to plunge in when I get back (third week of June); however, the mention of a special issue of _Minds and Machines_ devoted to the topic makes me think that I had at least declare my intentions now, lest I be left behind. What I have in mind is writing a brief (eg, 3000 word) "position paper" on the topic, with reference to the discussion so far mostly to give credit. But, as indicated, I can't get to it for a while. Is there any possibility of this, or does the timing wipe me out?

John Haugeland haugelan@unix.cis.pitt.edu

[Reply: The Symposium will continue, so there is still time for John Haugeland and others to join in. SH.]

Date: Fri, 8 May 92 15:58:43 EDT From: "Stevan Harnad"

Drew McDermott (mcdermott-drew@CS.YALE.EDU) wrote:

>dm> We're talking about whether semantic interpretability is part of the
>dm> *definition* of computer. For that to be the case, everything the
>dm> computer does must be semantically interpretable. Does it cease to be a
>dm> computer during the interludes when its behavior is not interpretable?

There is a systematic misunderstanding here. I proposed semantic interpretability as part of the definition of computation. A computer would then be a device that can implement arbitrary computations. That doesn't mean everything it does must be semantically interpretable. Uninterpretable states in a computer are no more problematic than idle or power-down states. What I suggest, though, is that if it was ONLY capable of uninterpretable states (or of only being idle or off), then it would not be a computer.

>dm> I assumed that your original claim was that a computer had to
>dm> correspond to an interpreted formal system (where, in the usual case,
>dm> the users supply the interpretation). But that's not what you meant...
>dm> now it's clear that neither you nor Pat is proposing anything of
>dm> this sort. Instead, you seem to agree with me that a computer is a
>dm> physical embodiment of a formal automaton, plus a kind of loose,

>dm> pragmatic, fault-prone correspondence between its states and various
>dm> world states. Given this agreement, let's simplify. Clearly, the
>dm> semantic interpretation is no part of the definition of computer. We
>dm> can identify computers without knowing what interpretation their users
>dm> place on them.

The interpretation of any particular computer implementing any particular computation is not part of my proposed definition of a computer. A computer is a physical system with the capacity to implement (many, approximately all) nontrivial computations (= INTERPRETABLE symbol systems), where "nontrivial" is a cryptographic complexity-based criterion.

>dm> The [videogame] "players and moves" mostly don't exist, of course,
>dm> since they include entities like King Koopa and Princess Toadstool. The
>dm> child playing the game thinks (sort of) that the pictures on the screen
>dm> refer to a certain universe. Or maybe they *constitute* a universe.
>dm> It's hard to be precise, but I hope by now vagueness doesn't bother
>dm> you. Of course, the engineer that wrote the game knows what's
>dm> *really* going on. The input signals refer to presses of control
>dm> buttons by the game player. Output signals refer to shapes on a
>dm> screen. But it would be absurd to say that the game player's version
>dm> of the semantics is only an illusion, and the real purpose of the
>dm> system is to map buttons pushes onto screen alterations.

Not that absurd, but never mind. There are certainly many levels of interpretation (virtual systems) in some computers implementing some programs. One virtual system need not have primacy over another one. My point is made if there is any systematic interpretability there at all.

We should keep it in mind that two semi-independent questions are under discussion here. The first has nothing to do with the mind. It just concerns what computers and computation are. The second concerns whether just a computer implementing a computer program can have a mind. The groundedness of the semantics of a symbol system relates to this second question. Computer video-games and their interpretations are hopelessly equivocal. They are just implemented squiggles and squoggles, of course, which are in turn interpretable as referring to bit-maps or to images of fictitious entities. But the fictitious entities are in OUR heads, and even the perception of the "entities" on the video-screen are mediated by our brains and their sensory apparatus. Without those, we have only squiggles and squoggles (or, in the case of the dedicated video system, hard-wired to its inputs and outputs, we have squiggles and squoggles married to buttons and bit-mapped CRT screens).

>dm> You're forgetting which side of the argument you're on. *I'm* arguing
>dm> that such interpretations are epiphenomenal. *You're* arguing that
>dm> the interpretation is the scaffolding supporting the computerhood of
>dm> the system.

You seem to be confusing the question of interpretability with the question of the groundedness of the interpretation. My criterion for computerhood is the capacity to implement arbitrarily many different (nontrivially) interpretable symbol systems. The interpretability of those systems is critical (in my view) to their being computational at all. Without interpretability you have random gibberish, uninterpretable in principle. But even interpretable (nonrandom) symbol systems are just gibberish unless we actually project an interpretation on them. This suggests that interpretability is not enough. If ANY kind of system (computational or not) is to be a viable candidate for implementing MENTAL states, then it cannot be merely interpretable; the interpretation has to be INTRINSIC to the system: it has to be grounded, autonomous, independent of whatever we do or don't project onto it.

Because of the grip of the hermeneutic circle, it is very hard, once we have projected an interpretation onto a system, to see it for what it really is (or isn't) on its own, independent of our interpretations. That's why I recommend de-interpreting candidate systems -- reducing them to the gibberish ("squiggles and squoggles") that they really are, to see what (if anything) is left to ground any meanings in. A pure symbol system (like some of the earlier overinterpreted chimpanzee "languages") could not survive this nonhermeneutic scrutiny. A TTT-scale robot could.

>sh> If you agree with Searle that computers can't be distinguished from
>sh> non-computers on the basis of interpretability, then I have to ask you
>sh> what (if anything) you DO think distinguishes computers from
>sh> non-computers?

>dm> I refer you to Chalmers. A brief summary: A system is a computer if
>dm> its physical states can be partitioned into classes that obey a
>dm> transition relation. -- Drew McDermott

I too think computers/computation can be distinguished from their (non-empty) complement, and perhaps by the elaboration of a criterion like that one. But this still leaves us miles apart on the question: "Ok, given we can objectively distinguish computers from noncomputers, what has this to do with the question of how to implement minds?"

Stevan Harnad

Date: Mon, 4 May 1992 10:31:25 +0200 From: Oded.Maler@irisa.fr (Oded Maler)

One outcome (at least for me) of the previous round of postings on the symbol-grounding problem (1990) was that I became aware of the fact that current computational models are not suitable for dealing with the phenomenon of computers interacting in real-time with the real world. Consequently, with several collaborators, I did some preliminary work on what we call "hybrid dynamical systems" which combine discrete state-transition dynamics with continuous change. This is technical work, and it is not supposed to solve the philosophical problems discussed here; I mention it just to show that such discussions, even if they don't seem to converge, might have some useful side-effects.

Now to the question of what is a computation. My current view is that computations are idealized abstract objects that are useful in describing the structure and the behavior of certain systems by focusing on the "informational" aspects of their dynamics rather on the "materialisic/energetic" aspects. This abstraction, not surprisingly, turns out to be useful in designing and analyzing certain systems such as synchronized switching devices, also known as general-purpose computers. It is sometimes also useful for analyzing the behavior of humans when they perform tasks such as adding numbers.

The question of why such a computational interpretation is more reasonable for some systems than for others is intriguing, and I don't know if a quantitative observer-independent borderline can be put. Even real airplanes do not necessarily fly, unless flying is a useful abstraction for us when we want to get to a conference - "you cannot participate in the same flight twice" (to rephrase what's-his-name, badly translated from Greek to Hebrew to English).

So I think the question will reduce to the two related problems: (1) What is "information"? -- because this seems to be the characterizing feature of computational dynamics. (2) The relations between things and their descriptions.

Oded Maler

From: Stevan Harnad (harnad@princeton.edu)

For me, flying is not just a "useful abstraction," it's something you really do, in the real air, otherwise you really fall. I agree with you that one of the problems here concerns the relation between things and their descriptions: The problem is when we confuse them! (And the concept of "information" alas seems just as subject to the problem of intrinsic versus derived meaning (i.e., groundedness) as computation is.)

Stevan Harnad

Date: Thu, 23 Apr 92 21:25:14 -0400 From: davism@turing.cs.nyu.edu (Martin Davis)

Stevan,

I've been watching the (real and virtual) stones flying in this discussion, amazed that none of the hermeneutic mirrors are broken. I had resolved to be safe and shut up. But here goes! I'm throwing, not stones, but rather, all caution to the wind.

Please forgive me, but this is what I really think: if and when brain function is reasonably well understood (and of course that includes understanding how consciousness works), this entire discussion will be seen as pointless, in much the same way that we now regard the battles that used to rage about the ether as pointless. In particular, I believe that the paradoxes of subjectivity ("How can I know that anyone other than me experiences redness?") will seem no more problematic than such equally compelling conundrums as: How can light waves possibly travel through empty space without a medium in which they can undulate? We (or rather our heirs) will know that other people experience redness because it will be known exactly what it is that happens in their brains and ours when redness is experienced. And then the objection that we cannot know that their experience is like ours, or even that they are experiencing anything, will just seem silly.

Whether a TT-passing computer is in any reasonable sense conscious of what it is doing is not a question we can hope to answer without understanding consciousness. If, for example, Dennett is right about consciousness, then I can perfectly well imagine that the answer could be "yes", since I can't see any reason why such mechanisms couldn't in principle be built into a computer program.

Martin Davis

Martin Davis (davism@turing.cs.nyu.edu) wrote:

md> if and when the brain function is reasonably well understood (and of md> course that includes understanding how consciousness works), this md> entire discussion will be seen as pointless... the paradoxes of md> subjectivity ("How can I know that anyone other than me experiences md> redness?") will seem no more problematic... [We] will know that other md> people experience redness because it will be known exactly what it is md> that happens in their brains and ours when redness is experienced. And md> then the objection that we cannot know that their experience is like md> ours, or even that they are experiencing anything, will just seem md> silly.

Martin,

You may be surprised to hear that this a perfectly respectable philosophical position (held, for example, by Paul Churchland and many others) -- although there are also MANY problems with it, likewise pointed out by many philosophers (notably, Tom Nagel) (and although the parenthetic phrase about "understanding how consciousness works" comes perilously close to begging the question).

But you will also be surprised to hear that this is not a philosophical discussion (at least not for me)! I'm not interested in what we will or won't be able to know for sure about mental states once we reach the Utopian scientific state of knowing everything there is to know about them empirically. I'm interested in how to GET to that Utopian state. And if it should be the case (as Searle and others have argued) that the symbolic road is NOT the one that leads there, I would want to know about that, wouldn't you? Perhaps this is the apt point to trot out (not for the first time in the symbol grounding discussion) the reflection of the historian J.B. Hexter on the value of negative criticism:

in an academic generation a little overaddicted to "politesse," it may be worth saying that violent destruction is not necessarily worthless and futile. Even though it leaves doubt about the right road for London, it helps if someone rips up, however violently, a "To London" sign on the Dover cliffs pointing south...

md> Whether a TT-passing computer is in any reasonable sense conscious of md> what it is doing is not a question we can hope to answer without md> understanding consciousness. If, for example, Dennett is right about md> consciousness, then I can perfectly well imagine that the answer could md> be "yes", since I can't see any reason why such mechanisms couldn't in md> principle be built into a computer program.

Yes, but if you have been following the discussion of the symbol grounding problem you should by now (I hope) have encountered reasons why such (purely symbolic) mechanisms would not be sufficient to implement mental states, and what in their stead (grounded TTT-passing robots) might be sufficient.

Stevan Harnad

Date: Fri, 8 May 92 17:04:08 EDT From: "Stevan Harnad"

From: dietrich@bingsuns.cc.binghamton.edu Eric Dietrich Subject: Re: Publishing the "What is Computation?" Symposium To: harnad@Princeton.EDU (Stevan Harnad)

> To all contributors to the "What is Computation?" Symposium:
>
> Please let me know which of you are (and are NOT) interested in
> publication. In the meanwhile, we can continue for a few more
> iterations before involking cloture. Perhaps the prospect of publication
> will change the style of interaction from this point on, perhaps not...

Stevan: I am interested in publication.

Eric Dietrich

Date: Fri 8 May 92 13:39:59-PDT From: Laurence Press

Dear Steven,

Have you got a publisher in mind for the book? If not, I am a consulting editor at Van Nostrand Reinhold, and would be happy to talk with them about it.

Larry Press

From: Stevan Harnad To: Larry Press

Dear Larry,

The context for the idea was actually a proposal from the Editor of Minds and Machines, James Fetzer, to publish it as a special issue of his journal. Thanks for your offer too, but unless we encounter problems in fitting it within the scope of a special journal issue, it looks as if we're already spoken for!

Best wishes,

Stevan Harnad

From: Pat Hayes (hayes@cs.stanford.edu) Date: Tue, 28 Apr 92 18:04:15 MDT

Searle's argument establishes nothing. If one is inclined to accept the idea that an implemented program might correctly be said to exhibit cognition, then the scenario Searle outlines - which we all agree to be fantastic, but for different reasons - suggests that there is an important difference between a computer running a program and the process of a human following through the steps of an algorithm, and if one were to achieve the former with a human as the computer then one would have a(n admittedly fantastic) situation, something akin to a fragmented personality. If one is inclined to reject that idea, Searle's scenario can be taken as further bolstering of that inclination, as many have noted.

I don't think the other-minds 'barrier' is really germane to the discussion, as it applies as much to other humans as to (hypothesised) artifical agents. I take it as observationally obvious that stones don't have minds, that (most) humans do, and that such things as cats and mice and perhaps some complex computational systems are best described as having partial, simple, or primitive minds. Somewhere between cats (say) and snails (say) the concept becomes sufficiently unclear as to probably be worthless. (This gradual deterioration of mentality is not a crisp phase transition, by the way, and I don't think that there is such a sharp division between mere biology or mechanism and real intensional thought.)

You wrote:

>
>sh> normally the only way to know whether or not a
>
>sh> system has a mind is to BE the system.

If one takes this stark a view of the other-minds question then it seems to me hard to avoid solipsism; and I may not be able to refute solipsism, but I'm not going to let anyone ELSE persuade me its true.

We can go on disagreeing for ever, but let me just say that I don't feel any sense of strain or cost to maintain my views when shown Searle's curious misunderstandings of computational ideas.

>
>ph> If we include (as we should) linguistic input, it seems clear that
>
>ph> structures and processes [underlying our capacity to categorize] will
>
>ph> be largely symbolic... vision and other perceptual modes involve
>
>ph> symbols from an early stage...
>
>
>sh> The only problem with "including" (as you put it) linguistic input is
>
>sh> that, without grounding, "linguistic input" is just meaningless
>
>sh> squiggles and squoggles. To suppose it is anything more is to beg the
>
>sh> main question at issue here.

Oh no, I have to profoundly disagree. The question is how formal symbols in a computational system might acquire meaning. But surely the words in the English sentences spoken to a machine by a human do not need to have their meaningfulness established in the same way. To take English spoken by humans - as opposed to formalisms used by machines - as having content surely does not beg any of the questions we are discussing.

>
>sh> To categorize is to sort the objects in the world, beginning with their
>
>sh> sensory projections

But surely by insisting on beginning thus, YOU are begging the question!

>
>sh> It is true that we can sort names and descriptions
>
>sh> too, but unless these are first grounded in the capacity to sort and
>
>sh> name the objects they refer to, based on their sensory projections,
>
>sh> "names and descriptions" are just symbolic gibberish...

Rhetoric again. But look at this carefully. Consider the word "everyone". What kind of ’sensory projection’ could provide the suitable ’grounding’ for the meaning of this? And try "whenever", "manager" or "unusual". Language is full of words whose meaning has no sensory connections at all.

>
>sh> But your point about the correspondence between the internal numerical
>
>sh> symbol for the length of an internal sequence can be made without
>
>sh> referring to the mental.

Yes. I did, actually.

>
>sh> There is certainly a correspondence there,
>
>sh> and the interpretation is certainly guaranteed by causality, but only
>
>sh> in a slightly more interesting sense than the interpretation that every
>
>sh> object can be taken to be saying of itself "Look, here I am!"

That particular example may not be very interesting, but the point it makes is rather more, since it is illustrative of a huge collection of computational phenomena throughout which interpretation is similarly guaranteed by causality.This was Brian Smith’s point: computation is, as it were, permeated by meanings causally linked to symbols.

>
>sh> Symbols that aspire to be the language of thought cannot just have a few
>
>sh> fixed connections to the world.

This raises a very interesting question. Let us suppose that you are basically right about the need for grounding to guarantee meaning. I believe you are, and have made similar points myself in my 'naive physics' papers, although I think that English can ground things quite successfully, so have more confidence in the TT than you do. But now, how much grounding does it take to sufficiently fix the meanings of the symbols of the formalisms? Surely not every symbol needs to have a direct perceptual accounting. We have all kinds of mechanisms for transferring meanings from one symbol to another, for example. But more fundamentally, beliefs relating several concepts represent mutual constraints on their interpretation which can serve to enforce some interpretations when others are fixed. This seems to be a central question: just how much attachment of the squiggles to their meanings can be done by axiomatic links to other squoggles?

>
>sh> The systematicity that is needed has to
>
>sh> have at least the full TT power of natural language -- and to be
>
>sh> grounded it needs TTT-scale robotic capacity.

Thats exactly the kind of assertion that I feel need not to be taken at face value.
>
>ph> What is different in having a machine that can run
>
>ph> algorithms from just being able to run algorithms? I take it as
>
>ph> obvious that something important is...
>
>
>sh> I think you're missing my point. The important thing is that the
>
>sh> algorithm be implemented mindlessly, not that it be implemented
>
>sh> mechanically (they amount to the same thing, for all practical
>
>sh> purposes). I could in principle teach a (cooperative) two-year old who
>
>sh> could not read or write to do rote, mechanical addition and
>
>sh> multiplication. I simply have him memorize the finite set of
>
>sh> meaningless symbols (0 - 9) and the small set of rules (if you see "1"
>
>sh> above "3" and are told to "add" give "4", etc.). I would then have a
>
>sh> little human calculator, implementing an algorithm, ...

No, thats exactly where I disagree. A human running consciously through rules, no matter how 'mindlessly', is not a computer implementing a program. They differ profoundly, not least for practical purposes. For example, you would need to work very hard on keeping a two-year-old's attention on such a task, but the issue of maintaining attention is not even coherent for a computer.

I know you find observations like this irrelevant to the point you are making - hence your quick "cooperative" to fend it off - but they are very relevant to the point I am making. I see an enormous, fundamental and crucial difference between your 'mindless' and 'mechanical'. The AI thesis refers to the latter, not the former. To identify them is to abandon the whole idea of a computer.

>
\>sh> Now let me tell you what WOULD be cheating: If any of what I had the
>
\>sh> child do was anything but SYNTACTIC, i.e., if it was anything other than
>
\>sh> the manipulation of symbols on the basis of rules that operate only on
>
\>sh> their (arbitrary) shapes: It would be cheating if the child (mirabile
>
\>sh> dictu) happened to know what "odd" and "even" meant, and some of the
>
\>sh> calculations drew on that knowledge instead of just on the mechanical
>
\>sh> algorithm I had taught him. But as long it's just mechanical syntax,
>
\>sh> performed mindlessly, it makes no difference whatsoever whether it is
>
\>sh> performed by a machine or stepped through (mechanically) by a person.

I disagree: I think it makes a fundamental difference, and to deny this is to deny that computation is real. But we are just beating our chests at one another again.

>
\>sh> Now if you want to appreciate the real grip of the hermeneutical circle,
>
\>sh> note how much easier it is to believe that an autonomous black box is
>
\>sh> "really" understanding numbers if it is a machine implementing an
>
\>sh> algorithm mechanically rather than an illiterate, non-numerate child,
>
\>sh> who is just playing a symbolic game at my behest.

Nah nah. You are just (re)in-stating the chinese room 'argument' AGAIN. And it still is convincing if you believe its conclusion, and not if you don't. It doesn't get anywhere.

>
\>sh> THAT's why you want to
>

>sh> disqualify the latter as a "real" implementation, despite the fact that
>
>sh> the same syntactic algorithm is being implemented in both cases, without
>
>sh> any relevant, nonarbitrary differences whatsoever.

No, I repeat: a human running through an algorithm does not constitute an IMPLEMENTATION of that algorithm. The difference is precisely what computer science is the study of: how machines can perform algorithms without human intervention. If you could get your two-year-old's body to IMPLEMENT addition algorithms, you would almost certainly be liable for criminal action.

>
>ph> Clearly if you insist that [reducing to gibberish]
>
>ph> can always be done to computer insides but not always to human
>
>ph> insides, then you are never going to see meaning in a machine.
>
>
>sh> I am sure that whatever is REALLY going on in the head can also be
>
>sh> deinterpreted, but you mustn't put the cart before the horse: You
>
>sh> cannot stipulate that, well then, all that's really going on in the
>
>sh> head is just symbol manipulation, for that is the hypothesis on trial
>
>sh> here!

Well, hang on. Surely if you concede that the head's machinations can be de-interpreted, then indeed you have conceded the point; because then it would follow that the head was performing operations which did not depend on the meanings of its internal states. That this is the point at issue does not make it illegal for me to have won the argument, you know. But maybe you did not mean to give that up so quickly. I will let you take that move back before claiming checkmate.

>
>sh> {Actually, there are two semi-independent hypotheses on trial: (1) Is
>
>sh> anything NOT just a computer doing computation? and, (2) Are minds just
>
>sh> computers doing computation? We agree, I think, that some things are
>
>sh> NOT computers doing computation, but you don't think the mind is one of
>
>sh> those noncomputational things whereas I do.]

Lets agree to dismiss (1). This Searlean thesis that everything is a computer is so damn silly that I take it simply as absurd. I don't feel any need to take it seriously since I have never seen a careful argument for it, but even if someone produces one, that will just amount to a reductio tollens disproof of one of its own assumptions.

>
>sh> I had recommended the exercise of deinterpreting the symbols so as to
>
>sh> short circuit the persuasive influence of those properties that are
>
>sh> merely byproducts of the interpretability of the symbols, to see
>
>sh> whether there's anything else left over. In a grounded TTT-scale robot
>
>sh> there certainly would be something left over, namely, the robotic
>
>sh> capacity to discriminate, categorize and manipulate the objects, events
>
>sh> and states of affairs that the symbols were about. Those would be there
>
>sh> even if the symbols were just gibberish to us. Hence they would be
>
>sh> grounding the interpretations independently of our mentalistic
>
>sh> projections.

OK. But what I don't follow is why you regard the conversational behavior of a successful passer of the TT clearly insufficient to attach meaning to its internal representations, while you find Searle's response to the "Robot Reply" quite unconvincing. If we are allowed to look inside the black box and de-interpret its innards in one case, why not also the other? Why is robotic capacity so magical in its grounding capacity but linguistic capacity, no matter how thorough, utterly unable to make symbols signify? And I don't believe the differences are that great, you see. I think much of what we all know is attached to the world through language. That may be what largely differentiates us from the apes: we have this incredible power to send meaning into one anothers minds.

Pat Hayes

(PS. The arithmetic example, while very simple, provides an interesting test for your hermeneutic intuitions. Take two different addition algorithms. One is the usual technique we all learned involving adding columns of numbers and carrying the tens, etc. . The other has a bag and a huge pile of pebbles and counts pebbles into the bag for each number, then shakes the bag and counts the pebbles out, and declares that to be the sum. A child might do that. Would you be more inclined to say that the second, pebble-counting child understood the concept of number? You can no doubt recognise the path I am leading you along.)

Date: Sun, 10 May 92 13:24:01 EDT From: "Stevan Harnad"

SYMBOLS CANNOT GROUND SYMBOLS

Pat Hayes (hayes@cs.stanford.edu) wrote:

>ph> I take it as observationally obvious [1] that stones don't have minds,
>ph> [2] that (most) humans do, and that such things as [3] cats and mice
>ph> and perhaps some [4] complex computational systems are [5] best
>ph> described as having partial, simple, or primitive minds... and I don't
>ph> think that there is such a sharp division between mere biology or
>ph> mechanism and real intensional thought.

Although I don't think this kind of "observation" is quite the same as other empirical observations, let me point out that one can readily agree with [1 - 3] and utterly disagree with [4], which suggests it might not all be so "obvious."

Let me also point out on exactly what MY judgment, at least, is based in these 4 cases. It is based purely on TTT-indistinguishability (note that I said TTT, i.e., total indistinguishability in robotic capacities, not merely TT, i.e., indistinguishability only in symbolic capacities).

(Although there is another potential criterion, TTTT (neural) indistinguishability, I am enough of a functionalist, and believe the robotic degrees of freedom are narrow enough, to make this further constraint supererogatory; besides, the TTTT is certainly not why or how we judge that other people and animals have minds.)

Animals do not pass the human TTT, but they come close enough. So would robots, making their way in the world (but, for methodological reasons, only if they passed the human TTT; we unfortunately do not know enough about animals' TTT capacities to be able to trust our judgments about animal-robots' TTT-indistinguishability from their biological counterparts: this is a serious problem for bottom-up robotics, which would naturally prefer to take on the amphioxus TTT before facing the human TTT!).

But you really begin to equivocate with [5]: "best described as having partial, simple, or primitive minds," because, you see, what makes this particular question (namely, the "other minds problem," pace Martin Davis) different from other empirical problems is that it is not merely a question of finding the "best description," for there also happens to be a FACT of the matter: There either IS somebody home in there, experiencing experiences, thinking thoughts, or NOT. And if not, then attributing a mind to it is simply FALSE, whether or not it is the "best description" (see Oded Maler's point about things vs. descriptions).

Nor is there a continuum from the mental to the nonmental (as there perhaps is from the living to the nonliving). There may be higher and lower alertness levels, there may be broader and narrower experiential repertoires of capacities, but the real issue is whether there is anybody home AT ALL, experiencing anything whatever, and that does indeed represent a "sharp division" -- though not necessarily between the biological and the nonbiological.

Now no one can know where that division really lies (except by being the candidate), but we can try to make some shrewd empirical inferences. Symbolic Functionalism ("thinking is just computation") was a natural first pass at it, but I, at least, think it has been shown to be insufficient because of the symbol grounding problem. Robotic Functionalism ("thinking is what goes on inside grounded TTT-scale robots") could be wrong too, of course, but until someone comes up with a principled reason why, I see no cause to worry about heading in that empirical direction.

>
>sh> normally the only way to know whether or not a
>
>sh> system has a mind is to BE the system.
>
>ph> If one takes this stark a view of the other-minds question then it
>ph> seems to me hard to avoid solipsism; and I may not be able to refute
>ph> solipsism, but I'm not going to let anyone ELSE persuade me its true.

For mind-modellers, the other-minds problem is not a metaphysical but a methodological problem. Abandoning computationalism certainly does NOT commit us to solipsism.

>ph> The question is how formal symbols in a computational system might
>ph> acquire meaning. But surely the words in the English sentences spoken
>ph> to a machine by a human do not need to have their meaningfulness
>ph> established in the same way. To take English spoken by humans - as
>ph> opposed to formalisms used by machines - as having content surely does
>ph> not beg any of the questions we are discussing.

There's no problem with the content of English for English speakers. The problem is with the content of English for a computer. English is grounded only in the heads of minds that understand what it means. Apart from that, it's just (systematically interpretable) squiggles and squoggles. The question is indeed how the squiggles and squoggles in a computer might acquire meaning -- and that certainly isn't by throwing still more ungrounded squiggles and squoggles at them...

>
>sh> To categorize is to sort the objects in the world,
>
>sh> beginning with their sensory projections
>
>ph> But surely by insisting on beginning thus, YOU are begging the question!

Not at all, I'm trying to answer it. If we start from the recognition that the symbols in a computer are ungrounded and need to be grounded, then one possible grounding hypothesis is that the requisite grounding comes from constraints exerted by symbols' physical connections to the analog structures and processes that pick out and interact with the the real-world objects that the symbols are about, on the basis of their sensorimotor projections. It seems to me that to attempt to ground systems other than from the sensory-bottom upward is to try to get off the ground by one's (symbolic?) bootstraps, or by clutching a (symbolic?) skyhook. I am, however, interested in rival grounding hypotheses, in particular, non-sensory ones, just as long as they are GROUNDING hypotheses and not just ways of letting the hermeneutics in by the back door (as in imagining that "natural language" can ground symbols).

>ph> Consider the word "everyone". What kind of "sensory projection" could
>ph> provide the suitable "grounding" for the meaning of this? And try
>ph> "whenever", "manager" or "unusual". Language is full of words whose
>ph> meaning has no sensory connections at all.

These objections to bottom-up sensory grounding have been raised by philosophers against the entire edifice of empiricism. I have attempted some replies to them elsewhere (e.g. Harnad 1992), but the short version of the reply is that sensory grounding cannot be investigated by armchair introspection on word meanings; it will only be understood through empirical attempts to design grounded systems. What can be said, however, is that most words need not be grounded directly. The symbol string "An X is a Y that is Z" is grounded as long as "Y" and "Z" are grounded, and their grounding can likewise be symbolic and indirect. The sensory grounding hypothesis is simply that eventually the symbolic descriptions can be cashed into terms whose referents can be pick out from their direct sensory projections.

"Everyone," for example, perhaps means "all people." "People," is in turn beginning to sound more like something we could pick out from sensory projections. Perhaps even the "all/not-all" distinction is ultimately a sensory one. But I'm just introspecting too now. The real answers will only come from studying and then modeling the mechanisms underlying our (TTT) capacity for discrimination, categorization and identification.

>ph> [There is] a huge collection of computational phenomena throughout
>ph> which interpretation is similarly guaranteed by causality [as in the
>ph> length of the internal string example]. This was Brian Smith's point:
>ph> computation is, as it were, permeated by meanings causally linked to
>ph> symbols.

And I'm not unsympathetic to that point; I just want to see it worked out and then scaled up to the TTT.

>
>sh> Symbols that aspire to be the language of thought
>
>sh> cannot just have a few fixed connections to the world.
>
>ph> Let us suppose that you are basically right about the need for
>ph> grounding to guarantee meaning. I believe you are, and have made
>ph> similar points myself in my "naive physics" papers, although I think
>ph> that English can ground things quite successfully, so I have more
>ph> confidence in the TT than you do. But now, how much grounding does it
>ph> take to sufficiently fix the meanings of the symbols of the formalisms?
>ph> Surely not every symbol needs to have a direct perceptual accounting.
>ph> We have all kinds of mechanisms for transferring meanings from one
>ph> symbol to another, for example.

These are empirical questions. I have no idea a priori how large a direct sensory basis or "kernel" a grounded TTT system requires (although I do suspect that the kernel will be provisional, approximate, and always undergoing revision whose consequences accordingly percolate throughout the entire system). But I am sure that "English" won't do it for you, because, until further notice, English is just systematically interpretable gibberish, and it's the interpretations that we're trying to ground!

Your phrase about "the need for grounding to guarantee meaning" also worries me, because it sounds as if grounding has merely a confirmatory function: "The meanings are already in the squiggles and squoggles, of course; we just need the robotic evidence to convince the sceptics." Well I think the meaning will be in the grounding, which is why I believe most of the actual physical structures and processes involved will be analog rather than computational.

>ph> But more fundamentally, beliefs relating several concepts represent
>ph> mutual constraints on their interpretation which can serve to enforce
>ph> some interpretations when others are fixed. This seems to be a central
>ph> question: just how much attachment of the squiggles to their meanings
>ph> can be done by axiomatic links to other squoggles?

The constraints you speak of are all syntactic. What they give you (if they are set up properly) is the coherent semantic INTERPRETABILITY that makes a symbol system a symbol system in the first place. The GROUNDING of that interpration must come from elsewhere. Otherwise it's just the self-confirmatory hermeneutic circle again.

>ph> A human running consciously through rules, no matter how "mindlessly,"
>ph> is not a computer implementing a program. They differ profoundly, not
>ph> least for practical purposes. For example, you would need to work very
>ph> hard on keeping a two-year-old's attention on such a task, but the
>ph> issue of maintaining attention is not even coherent for a computer.
>ph>
>ph> I see an enormous, fundamental and crucial difference between your
>ph> "mindless" and "mechanical." The AI thesis refers to the latter, not
>ph> the former. To identify them is to abandon the whole idea of a
>ph> computer... to deny this is to deny that computation is real.
>ph>
>ph> a human running through an algorithm does not constitute
>ph> an IMPLEMENTATION of that algorithm. The difference is precisely what
>ph> computer science is the study of: how machines can perform algorithms
>ph> without human intervention.

I suppose that if computer science were just the study of hardwares for implementing programs then you would have a point (at least about what computer scientists are interested in). But isn't a lot of computer science implementation-independent (software)? If someone writes a program for factoring polynomials, I don't think he cares if it's executed by a machine or an undergraduate. Usually such a program is written at a lower level than the one at which an undergraduate would want to work at, but the undergraduate COULD work at that lower level. I think the programmer would have to agree that anyone or anything following the syntactic steps his program specified would be "executing" his program, even if he wanted to reserve "implementing" it for the kind of mechanical implementation you are stressing.

I am not implying that designing mechanical devices that can mechanically implement programs is not an extremely important achievement; I just think the implementation-independence of the programming level renders all these hardware-related matters moot or irrelevant for present purposes. If I had to pick the two main contributions of computer science, they would be (1) showing how much you could accomplish with just syntax, and (2) building devices that were governed mechanically by syntax; most of the action now is in (1) precisely because it's independent of (2).

Let me try to put it another way: Prima facie, computer-hardware-science is a branch of engineering; it has nothing to do with the mind. What principle of hardware science could possibly underwrite the following distinction: If a program is executed by a machine, it has a critical property that it will lack if the very same program is executed by a person. You keep stressing that this distinction is critical for what counts as a true "implementation" of a program. So let's try to set trivial semantics aside and speak merely of the program's being "executed" rather than "implemented." What is there in COMPUTER SCIENCE that implies that mechanical execution will have any relevant and radically different properties from the human execution of the very same program (on the very same I/O)?

Now I realize the case of mind-implementation is unique, so perhaps you could give some suggestive examples of analogous radical differences between mechanical and human implementations of the same programs in other domains, just to set my intuitions.

>ph> Surely if you concede that the head's machinations can be
>ph> de-interpreted, then indeed you have conceded the point; because then
>ph> it would follow that the head was performing operations which did not
>ph> depend on the meanings of its internal states.

Not at all. All that follows from my very willing concession is that one can de-interpret any kind of a system at all, whether it is purely symbolic or not. WHATEVER is going on inside a grounded TTT-scale robot (you seem to be able to imagine only computation going on in there, but I can think of plenty more), whether we know its interpretation or not, those inner structures and processes (whatever they are) retain their systematic relation to the objects, events and states of affairs in the world that (unbeknownst to us, because de-interpreted) they are interpretable as being about. Why? Because those goings-on inside the head would be governed by the system's robotic capacity to discriminate, categorize, manipulate and discourse (in gibberish, if we don't happen to know the code) about the world TTT-indistinguishably from the way we do. In other words, they would be grounded.

>ph> what I don't follow is why you regard the conversational behavior of a
>ph> successful passer of the TT clearly insufficient to attach meaning to
>ph> its internal representations, while you find Searle's response to the
>ph> "Robot Reply" quite unconvincing. If we are allowed to look inside the
>ph> black box and de-interpret its innards in one case, why not also the
>ph> other? Why is robotic capacity so magical in its grounding capacity but
>ph> linguistic capacity, no matter how thorough, utterly unable to make
>ph> symbols signify? And I don't believe the differences are that great,
>ph> you see. I think much of what we all know is attached to the world
>ph> through language. That may be what largely differentiates us from the
>ph> apes: we have this incredible power to send meaning into one another's
>ph> minds.

WE do, but, until further notice, computers don't -- or rather, their capacity to do so (bidirectionally, as opposed to unidirectionally) is on trial here. To get meaning from discourse (as we certainly do), the meanings in our heads have to be grounded. Otherwise all that can be gotten from discourse is syntax. This is why the TT alone is inadequate: because it's all just symbols; nothing to ground the meanings of meaningless squiggles in except still more, meaningless squiggles.

I don’t find Searle’s response to the Robot Reply unconvincing, I find the Robot Reply unconvincing. It merely amounted to pointing out to Searle that people could do more than just write letters. So Searle said, quite reasonably, fine, add on those extra things and I still won’t understand Chinese. He was right, because the objection was wrong. It’s not a matter of symbol crunching PLUS some add-on peripherals, where the symbol-crunching is the real bearer of the meaning. That’s just as equivocal as symbol crunching alone.

No, my reply to Searle (which in Harnad 1989 I carefully dubbed the "Robotic Functionalist Reply," to dissociate it from the Robot Reply) explicitly changed the test from the TT to the TTT and accordingly changed the mental property in question from "understanding Chinese" to "seeing" in order to point out that even transduction is immune to Searle’s argument.

To put it in the briefest possible terms: Symbols alone will not suffice to ground symbols, and language is just symbols (except in the heads of grounded systems -- which neither books nor computers are).

>ph> The arithmetic example, while very simple, provides an interesting
>ph> test for your hermeneutic intuitions. Take two different addition
>ph> algorithms. One is the usual technique we all learned involving adding
>ph> columns of numbers and carrying the tens, etc. The other has a bag
>ph> and a huge pile of pebbles and counts pebbles into the bag for each
>ph> number, then shakes the bag and counts the pebbles out, and declares
>ph> that to be the sum. A child might do that. Would you be more inclined
>ph> to say that the second, pebble-counting child understood the concept of
>ph> number? You can no doubt recognise the path I am leading you along.
>ph>
>ph> Pat Hayes

You’ve changed the example a bit by having the child know how to count (i.e., able to attach a name to an object, namely, a quantity); this is beginning to leave behind the point, which was that we only wanted the child to do syntax, slavishly and without understanding, the way a computer does.

But, fair enough, if what you are talking about is comparing two different algorithms for addition, one involving the manipulation of numerals and the other the manipulation of pebbles (again on the assumption that the child does not have any idea what all this means), then I have no problem with this: Either way, the child doesn’t understand what he’s doing.

If you have two I/O equivalent algorithms you have weak equivalence (that’s what the TT is based on). The stronger equivalence (I called it Turing Equivalence, but you indicated [in Hayes et al 1992] that that was the wrong term) requires two implementations of the same algorithm, both equivalent state for state. The latter was the equivalence Searle was considering, and even with this strong form of equivalence there’s no understanding.

Your point is not, I take it, about how one goes about TEACHING arithmetic to a child, or about what a child might figure out from a task like this -- for that’s just as irrelevant as the question of whether or not Searle might actually learn a few things about Chinese in the Chinese room. All such considerations beg the question, just as any verbal instruction to either the child or Searle (about anything except the syntactic rules to be followed) would beg the question.

Stevan Harnad

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on the Virtual Mind. Minds and Machines (in press)

Date: Wed, 29 Apr 92 12:45:24 EST From: Ross Buck

I have been aware of the symbol grounding discussion and discussion re computation, but have admittedly not been keeping up in great detail, and if this is too off-the-wall please ignore it. However, I have a perspective that might be relevant. I have made a distinction between special-purpose processing systems (SPPSs) structured by evolution and general-purpose processing systems (GPPSs) stuctured during the course of ontogeny. Perceptual systems are an example of the former and classical conditioning, instrumental learning, and higher-order cognitive processes examples of the latter. A Gibsonian view of perceptual systems suggests that perception is DIRECT in that there are evolved compatibilities between events and the experience of the perceiver. The experience is not a symbol of the event, it is a SIGN, which in my definition bears a natural relationship with its referent. A sign so defined may be what you call a grounded symbol: I’m not sure.

I would argue that GPPSs are computers but SPPSs are not. SPPSs are a result of evolution and are a defining characteristic of living systems. Computers are not living systems, but living systems have incorporated computers, in the form of GPPSs.

References:

Buck, R. (1985) Prime theory: A general view of motivation and emotion. Psych. Review.

Buck, R. (1988) Human Motivation and Emotion. 2nd. Ed. New York: Wiley.

==

From: Stevan Harnad

[The above contribution was returned to the author several weeks ago with the following comment; the rest was subsequently added by the author in response.]

Ross, the two kinds of systems you mention sound like they are worth distinguishing, but you have not given the specifics of what goes on inside them: Is it computation (symbol manipulation) or something else (and if something else, then what)? So far you have sorted two kinds of package, but what we are discussing is the contents. I will only be able to post your contribution to the SG list as a whole if it takes up the substance of the discussion. (The Peircian terminology does not help either, unless you specify a mechanism.) -- Stevan Harnad

==

It is something else, namely reproduction. Whereas GPPSs inherently are computers, SPPSs are designed to reproduce events unaltered, for all practical purposes. The analogy would be with reproduction devices like audio/video communication systems, rather than computers. The old Shannon-Weaver analysis comes to mind. More specifically:

Living things differ from computers in that the former have an inherent purpose: the maintenance of the DNA molecule. This involves maintaining the temperature, energy, and chemical (TEC) balances necessary for the DNA molecule to exist and replicate. The success of living things in this regard is demonstrated by the fact that the TEC balances existing within our bodies are roughly comparable to those of the primordial seas in which the DNA molecule first spontaneously generated. To maintain these balances early life forms evolved perceptual systems to monitor both the external environment and the internal, bodily environment for TEC resources and requirements; and response systems to act accordingly: to approach TEC states that favor DNA existence and replication and to avoid TEC states that endanger them. In the process, there evolved basic systems of metabolism (including the oxygen-burning metabolism of animals and the photosynthesis employed by plants which shaped the atmosphere of the early earth), complex eukaryotic cells, sexual reproduction, multicelled creatures, social organization, etc. By far the largest span of time of life on the earth has involved the cobbling together via evolution of these basic systems. Each system, in a teleonomic (as opposed to teleological) process, evolved to serve a specific function: for this reason I prefer to call them "special-purpose processing systems" (SPPSs).

One might argue that these systems involve computation: in a sense any process that involves information transfer might be defined as involving computation. I suggest that the term computation be reserved for systems involving information processing, and that systems designed around information transfer are fundamentally distinct: they are recording devices rather than computing devices. Recording systems have inherent "meaning:" the nature of the event being recorded. From the point of view of the DNA molecule it is critical that the information received by the perceptual systems regarding TEC events is accurate: that it matches in critical respects the actual TEC events. If this is not the case, that DNA molecule is unlikely to survive. The result is the evolution of a perceptual system along the lines of Gibsonian theory: compatibilities between the critical TEC events and the recording qualities of the system evolve naturally and inevitably, so that the organism gains veridical access to certain events in both the external terrestrial environment (including the activities of other organisms) and the internal bodily environment (the latter in the form of motivational-emotional states: in complex creatures who know THAT they have these states they constitute affects, or desires and feelings).

I term the elements by which information is transferred in SPPSs "signs" rather than symbols. This is admittedly a Pierceian term, but I do not wish to imply a Pierceian definition. Whereas symbols have arbitrary relationships with their referents, the relationship between the sign and the referent is natural. The living organism has access to important TEC events via signs of those events incorporated in the perceptual system: the photon excites a rod in the retina, which in turn excites a sensory neuron in the optic nerve, and so on to the visual cortex. Even though information is altered in form, the system is designed so that the meaning of the information--its relationship to the TEC events--is maintained: the sign of the event is passed up the line altered in form but not meaning. (Is this what you mean by a grounded symbol system, Stevan?)

The evolution of SPPSs took a long time: roughly three billion of the 3.8 billion year old story of life on earth. In the process, the environment of the earth was transformed: the oxygen atmosphere and ozone layer for example are products of life. Very early on, however, it became useful for creatures to process information as well as merely receive it: to associate one event with another, as in Pavlovian conditioning; to approach events associated with beneficial TEC outcomes (i.e., positive incentives) and avoid negative events, etc. This requires generalpurpose processing systems (GPPSs) that are structured by experience during the course of ontogeny: computing systems. In human beings, the range and power of such systems has been greatly increased by language.

Thus living things are not computers, but they have come to employ computing devices in adapting to the terrestrial environment. But the fundamental teleonomic goal of living systems--the meaning of life, as it were--is to maintain the TEC balances necessary for the existence of the DNA molecule. Ironically, the activities of human beings made possible by the power of the GPPSs and language have destroyed these balances beyond redemption for many species, and placed in jeopardy the future of life on the earth.

Ross Buck

From: Kentridge Date: Fri, 8 May 92 15:45:29 BST

[Here is] something for the computation discussion perhaps (I have missed a bit after being away - I hope I’m not retreading old ground or too completely off the mark for the current state of the discussion).

Dynamic properties of computational systems.

The discussion of the nature of computation has reached the issue of symbol interpretability just as previous discussions of Searle’s Chinese Room problem did. While I would not deny the importance of issues of symbol interpretation when considering adaptive intelligence I think one of the most interesting questions raised by Searle was "what is special about wetware?". I wish to consider an allied question "what is special about physical systems which compute?". In order to understand computation and intelligence I think we need to address both symbolic and physical issues in parallel. Perhaps some consideration of the physics of computation might resolve some of the paradoxes that the current symbolic discussion is producing.

There are two basic classes of dynamic behaviour in physical systems - attraction to equilibrium and attraction to chaos. I will consider the effects of introducing a signal which contains some information on which we wish the system to perform computation for both classes.

When we perturb an equilibrium by introducing a signal into it its response is to settle back into one of a finite number of stable equilibrium states. The transition from initial state via the perturbed state to a resultant, possible new, equilibrium state is entirely deterministic and predictable. Once the final state has been reached, however, the precise nature of the perturbing signal is lost. Stable states of the system do not preserve any information on the history of the signals which drove them there. Such systems can only perform trivial computation because of their limited memory - we can conceive of them as Turing machines with very very short tapes. In chaotic systems we fare no better, but for opposite reasons. In a chaotic system each different perturbation introduced to a system in a given starting state will produce a unique resulting behaviour (even if two perturbations push the system onto the same chaotic attractor the resulting orbits will never meet); the system has infinite memory. The problem, however, is that the transition between initial, perurbed and resultant states in unpredictable. The chaotic system is like a Turing machine with a very unreliable automaton.

One characteristic of systems that do computation then is that they are neither chaotic nor equilibrium systems, they can, however, be at the boundary or phase transition between these two regimes. In such systems distinct effects of perturbations can last arbitrarily long (but not infinite) times and transitions between states are at least probabilistically predictable. The notion that computation only occurs at phase transitions has received experimental support from studies of cellular automata (e.g. Packard, 1987) and theoretical support from analysis of the informational properties of the dynamics of phase transitions in terms of the relationship between complexity and entropy (Crutchfield and Young, 1991). Analysis of the dynamics of simulations of physiologically plausible models of cerebral cortex suggests that cortex may be well suited to being maintained in a critical dynamic state between equilibrium and chaos (Kentridge, 1991) in which computation can take place.

There are a few conclusions I would like to draw from this. First, in this scheme of things computation per se is defined solely in terms of the relationship between the complexity (the number of types of structural regularites which are needed to describe a set of data) and the entropy of data produced by a system. Processes which produce a high ratio of complexity to entropy are ones which are capable of computation. Second, as a consequence of this, everything is not a computer doing computation. Third, computation is only interpretable in terms of the the regularities that are used in the definition of complexity - if there is a correspondence between those regularities and the rest of the world then we may recognise the computation as being useful.

I hope this is of some help to someone! It seems to me at least that a physical definition of computation allows us to recognise systems as performing computation even if we can’t interpret computation. It also emphasizes that there is an important relationship between the hardware on which computation occurs and the nature of interpretable computation.

Caveat: I’m really a physiological psychologist so reference to the following sources is recommended (well the first two at least!).

References.

Packard, N.H. (1987) Adaptation towards the edge of chaos. In J.A.S. Kelso, A.J. Mandell and M.F. Schlesinger (Eds.) Dynamic patterns in complex systems. Singapore: World Scientific.

Crutchfield, J.P. and Young, K. (1991) Computation at the onset of chaos. In W.H. Zurek (Ed.) Complexity, entropy and the physics of information. (Proceeding of the Santa Fe Institute Studies in the Sciences of Complexity Volume 8.) Redwood City, CA.: Addison-Wesley.

Kentridge, R.W. (1991) Weak chaos and self-organisation in simple cortical network models. Eur. J. Neurosci. S4 73.

Robert Kentridge

Date: Sun, 10 May 92 19:29:44 EDT From: "Stevan Harnad"

Date: Fri, 8 May 92 17:58:50 PDT From: Dr Michael G Dyer Subject: Turing Machine - Brain Continuum

Here's another thought experiment for all. Imagine a continuum C: At one end is a physical brain B, capable of passing the Turing Test (or have it pass the TTT by symbols on parts of its tape controlling a robot, whichever you want).

At the other end of the continuum C is a Turing Tape T that is SO LONG and has so many "squiggles" that it models B at the molecular level. That is, for every mRNA twist/fold and protein etc. produced, there is a corresponding (huge) set of squiggles that encode their state. Transformations of squiggles etc. encode the molecular dynamics (and thus also the neural dynamics).

Notice that I've tried to remove the granularity issue (i.e. "big" symbols of traditional AI versus smaller distributed connectionist "subsymbols") by picking an extremely low (i.e. molecular) level of granularity.

Both T and B have identical behavior (wrt any TT or TTT scenarios you want to devise) so the behavior is also NOT the issue -- both T and B *act* intentional.

The strong AI people would (I assume) claim that both T and B have "intelligence", "consciousness" "intentionality", etc.

Searle (I assume) would claim that B has consciousness/intentionality but T does NOT.

Harnad (I assume) would claim that both T and B have consciousness only when controling the robot (i.e. TTT case) and both do NOT have it when "disembodied" (i.e. only passing the TT).

First, let's deal with Searle vs Strong AI. We do this by slowly moving along this continuum (of models), from B-to-T (or T-to-B). To move from B-to-T we replace segments (either scattered or contiguous, I don't think it matters) of B brain tissue with smaller Turing Machines Ti where each Ti performs some equivalent function performed by some subpart of B.

To move from T-to-B we replace bunches of squiggles on the tape with real cells, (or subcells, or cell assemblies, etc.).

The continuum might be viewed better as some kind of lattice, with the models in the middle being hybrid brains with different mixtures of cellular subsystems vs Turing Machines. For example, one hybrid is where EVERY OTHER B-neuron (with its dendrites/axons) is modeled by a separate Turing Machine, so the hybrid is a scattered mix of 50% real neurons and 50% Turing Machine simulations, all linked up.

(Turing Machines-to-cellular INTERFACES are the trickiest part. There are probably many ways of doing this. In this thought experiment I think it’s ok to scale Turing Machines to a very small size (i.e. super nontechnology) so they can be more easily interfaced with dendrites (and operate even within a cell). But the main requirement is that a real neuron or protein’s dynamics cause a corresponding representation to be placed on a Turing tape at the right time.)

In any case, ALL points on continuum C maintain the behavioral correspondence so that the behavior (for passing the TT or TTT) is the same.

Now, it seems to me that Searle is going to have a hard time determining when "consciousness" or "intentionality" appears, as one moves from T toward B. It’s clear that he will be happy with B and unhappy with T but what about all of the possibilities inbetween?

Now let’s create a new continuum C’ along the "sensory embodiment" dimension by extending C along this dimension. To do this we start out with both B and T controlling a complete robot, with hands/legs/mouth, eyes/ears/skin-sensors.

As we move along C’, we slowly remove these sensors/effectors. E.g., if there are 1million sensors/effectors, we cut them off, bit by bit, and leave only nerve "stumps" (in the B case) or in the T case, we cut the wires that allows a set of "squiggles" on a Turing tape to control the robot (or those wires that take sensory input and place "squiggles" on some "sensory" portion of the tape). We do this also for all the hybrid brain/machines in the middle of the continuum C. So we have now an "intentionality plane IP" of models CxC’. How you assign "intentionality/consciousness" labels to points/regions on this plane will then say something about your intuitions concerning consciousness.

Strong AI appears to be quite "liberal" -- i.e. assigning "intentionality" to the entire plane (since ALL points on IP demonstrate the same intentional BEHAVIOR).

Does Searle only assign intentionality to point B or does he accept intentionality at other points/regions ???

I’m not sure where Harnad assigns intentionality along C’. Will just ONE photo-sensitive cell be enough for "embodiment"? How many sensors/effectors are needed along continuum C’ before intentionality/consciousness appears/disappears for him? (Stevan, perhaps you can enlighten us all on this.)

Both Searle and Harnad just can’t accept that a TM (all those "mindless squiggles") could have a MIND. But to anyone accepting MIND as the ORGANIZATION of physical systems, our Turing Machine T has all the organization needed (with an admittedly tight bottleneck of just enough causal apparatus to have this organization direct the dynamics of the Turing tape read/write/move actions).

But is it any more of an amazing fact that "all those meaningless squiggles" create a Mind than the (equally) amazing fact that "all those mindless neurons" create a Mind? We’re simply USED to seeing brains show "mindfulness". We are not (yet) used to Turing-class machines showing much mindfulness.

Michael G. Dyer

From: Stevan Harnad

Fortunately, my reply to Mike can be very short: Real neurons don't just implement computations, and symbolically simulated neurons are not neurons. Set up a continuum from a real furnace heating (or a real plane flying, or a real planetary system, moving) to a computational simulation of the same and tell me where the real heating (flying, moving) starts/stops. It's at the same point (namely, the stepping-off point from the analog world to its symbolic simulation) that a real TTT-passing robot (with its real robot-brain) and its computationally simulated counterpart part paths insofar as really having a mind is concerned.

Stevan Harnad

Date: Fri, 8 May 92 10:18:00 HST From: Herbert Roitblat

I printed our various communications on this issue and it came to 125 pages. I think that we might want to summarize contributions rather than simply clean up what has been said.

I will contribute.

Herb Roitblat

From: Brian C Smith Date: Sat, 9 May 1992 15:54:29 PDT

I've enjoyed this discussion, but would very strongly like to argue against publishing it. Instead, I'd like to support John Haugeland's (implicit) suggestion.

For my money, there are two reasons against publishing.

First, I'm not convinced it would be interesting enough, per page. It is one thing to be part of such discussions -- or even to read them, a little bit each day, as they unfold. It has something of the structure of a conversation. It doesn't hurt that many of us know each other. Sitting down with the entire opus, as an outsider, is quite another thing. Last night I reread about a month's worth in paper form, imagining I were holding a journal in my hand -- and it didn't take. It just doesn't read like professional prose. This is not a criticism of anyone. It's just that the genre of e-mail discussion and the genre of referreed journal are different. Excellent one needn't make excellent the other.

More seriously, there has been no attempt to keep the format objective. Stevan has thoroughly mixed moderating and participating, making the result a public correspondence of his, more than a balanced group discussion. It is not just that he has contributed the most (50% of the total, four times as much as the nearest competitor [1]). It is more subtle things -- such as that, for example, a number of contributions [e.g. Sereno, Moody, Dambrosio, some of Hayes] only appeared embedded within his replies; others [e.g. Myers] only after being preceded with quite normative introduction. You don't need fancy analysis to see how much these things can skew the overall orientation.

Since it is Stevan's list, he is free to do this, and we are free to participate as we chose (though I must say that these things have limited my own participation quite a lot). I assume this is all as he intended. But it would be a very different thing to publish the result as in any sense a general discussion. Certainly to pose it as a general discussison that Stevan has merely would be quite a misrepresentation.

On the other hand, the topic is clearly of wider interest. So instead I suggest that we adopt John Haugeland's suggestion -- and that each of us write a 3000-5000 word brief or position paper on the question, and these be collected together and published. We can draw intellectually on the discussion to date -- but it would also give us a chance to distill what we've learned into punchier, more targeted form.

Brian

P.S. The prospect of publishing e-mail discussions clearly raises all kinds of complex issues -- about genre, the role of editing, applicable standards and authority, models of public debate, etc. I've just been focusing on one: of maintaining appropriate detachment between the roles of moderating and passionate participation. But many others deserve thinking through as well.

[1] 1057 lines out of 2109 counted between April 17 and May 8; Pay Hayes was second with 277.

From: Stevan Harnad

Well, I guess this calls for some sort of a reply from me:

(a) This electronic symposium began informally, with some cross-posting to a small group of individuals (about a dozen); only later did I begin posting it to the entire Symbol Grounding Discussion List (several hundred, which I have moderated for four years), with pointers to the earlier discussion, electronically retrievable by anonymous ftp.

(b) The expressions of interest in publication (one from James Fetzer, editor of Minds and Machines, about the possibility of publishing some version of the symposium as a special issue of his journal, and one from Laurence Press, consulting editor for Van Nostrand Rheinhold, expressing interest in publishing it as a book) came still later.

(c) No one had initially expected the symposium to reach the scope it did, nor to draw in as many participants as it has so far. In the very beginning I cross-posted the texts annotated with my comments, but once it became clear that the scale of participation was much larger than anticipated, I switched to posting all texts directly, with comments (my own and others') following separately, skywriting-style.

(d) In the published version (if there is to be one), all texts, including the earliest ones, would appear as wholes, with comments (and quotes) following separately. This is how we did it in editing and formatting the shorter "Virtual Mind" Symposium (Hayes et al. 1992) under similar circumstances.

(e) In moderating the symposium, I have posted all contributions I received in toto (with two exceptions, one that I rejected as irrelevant to the discussion, and one [from Ross Buck] that I first returned for some clarification; the revised version was subsequently posted).

(f) Mike Dyer (sic), with whom I have had long and stimulating exchanges in past years on the Symbol Grounding Discussion Group, entered this discussion of "What is Computation?" on our old theme, which concerns whether a computer can have a mind, rather than what a computer is. Since our respective views on this theme, which I think we have rather run into the ground, had already appeared in print (Dyer 1990, Harnad 1990), I hoped to head a re-enactment off of them at the pass. As it happens, both themes have now taken on a life of their own in this discussion.

(g) It is embarassing that I have contributed more to the symposium than others have (and the proportions could certainly be adjusted if it were published) but I must point out that this imbalance is not because others were not able -- indeed encouraged -- to contribute. Some (like Pat Hayes and Drew McDermott) availed themselves of the opportunity fully, others did not.

(h) There is no necessity at all that I, as the moderator of the symposium, be the editor of the published version, indeed I would be more than happy to cede this role to someone else.

(i) Regarding refereeing: James Fetzer indicated clearly that if it appeared in his journal, the published version would first be subject to peer review.

(j) I do wish to register disagreement with Brian Smith on one point, however: I would strongly favor publishing it as a symposium, one that preserves as much as possible of the real-time interactive flavor of this remarkable new medium of communication ("scholarly skywriting"). In reading over the unedited transcript as an "outsider," as Brian did, it is unavoidable that one’s evaluation is influenced by the fact that elements of the back-and-forth discussion are not all that congenial to one’s own point of view. The remedy for this is not to turn it into a series of noninteractive position papers, but to launch into more interactive participation. Afterward, editing and peer review can take care of making the symposium into a balanced, integrated, publishable final draft.

(k) Since I posted the two possibilities of publication, we have heard affirmatively about publication from (1) Dave Chalmers and (2) Eric Dietrich. I (3) favor publication too. We have now heard from (4) Herb Roitblat and (5) Brian Smith (whose view is seconded below by (X) John Haugeland, who has, however, not yet contributed to the symposium). How do the other 19 of the 24 who have so far contributed to the symposium feel about publication, and whether it should be in the form of an interactive symposium or a series of position papers?

(6) Frank Boyle (7) Ross Buck (8) John M Carroll (9) Jeff Dalton (10) Bruce Dambrosio (11) Martin Davis (12) Michael G Dyer (13) Ronald L Chrisley (14) Gary Hatfield (15) Pat Hayes (16) Robert Kentridge (17) Joe Lammens (18) Oded Maler (19) Drew McDermott (20) Todd Moody (21) John Searle (22) Marty Sereno (23) Tim Smithers (24) Richard Yee

Stevan Harnad

Dyer, M. G. (1990) Intentionality and Computationalism: Minds, Machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence, Vol. 2, No. 4.

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on the Virtual Mind. Minds and Machines [in press; published version of electronic "skywriting" symposium]

Harnad, S. (1990) Lost in the hermeneutic hall of mirrors. Invited Commentary on: Michael Dyer: Minds, Machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence 2: 321 - 327.

Date: Sat, 9 May 92 22:06:44 -0400 From: "John C. Haugeland" To: cantwell@parc.xerox.com, harnad@clarity, hayes@cs.stanford.edu Subject: Publishing

Brian, Pat, and Steve:

As usual, Brian knows me better than I know myself. I didn't realize that I was making an implicit proposal, or even that I wanted to. But now that I think about it, I do -- just as Brian said, and for just those reasons. Modest position papers, informed by the discussion so far, but not directly drawn from it, seem like much better candidates for publishable prose.

John Haugeland

Date: Sat, 09 May 92 17:51:40 ADT From: Lev Goldfarb

Oded Maler (Oded.Maler@irisa.fr) wrote:

om> Now to the question of what is a computation. My current view is that om> computations are idealized abstract objects that are useful in om> describing the structure and the behavior of certain systems by om> focusing on the "informational" aspects of their dynamics rather on om> the "materialisic/energetic" aspects.

Let me try to attempt a more formal definition:

Computation is a finite or infinite sequence of transformations performed on "symbolic" objects.

One can add that an "interesting" computation captures in some (which?) form the dynamics of some meaningful (to whom?) processes. It appears that the question marks cannot be removed without the participation of some intelligent (understood very broadly) entity that can interpret some sequences of transformations as meaningful.

Lev Goldfarb

Date: Sun, 10 May 92 15:33:41 -0400 From: davism@turing.cs.nyu.edu (Martin Davis) Subject: TTT ?

Stevan Harnad (harnad@clarity.princeton.edu) wrote:

>
\>sh> You may be surprised to hear that this a perfectly respectable
>
\>sh> philosophical position (held, for example, by Paul Churchland and
>
\>sh> many others)

Not so surprised. I've been a fan of Pat Churchland (I've never met Paul) and regard her work as being remarkably sensible. (By the way, I've held my present views for many years; I first formulated them explicitly in discussions with Hilary Putnam during the years we worked together ~1958.)

>
\>sh> (and although the parenthetic phrase about "understanding how
>
\>sh> consciousness works" comes perilously close to begging the
>
\>sh> question).

I carefully said "IF AND WHEN the brain function is reasonably well understood (and of course that includes understanding how consciousness works)". Of course, to believe that such understanding is likely to come and that with it will come understanding of "mind" and in just what sense "someone is home" up there, is to have a definite stance (I would once have said "materialistic") about such matters. But what question am I begging, or coming "perilously close" to so doing?

>
\>sh> But you will also be surprised to hear that this is not a
>
\>sh> philosophical discussion (at least not for me)! I'm not
>
\>sh> interested in what we will or won't be able to know for sure
>
\>sh> about mental states once we reach the Utopian scientific
>
\>sh> state of knowing everything there is to know about them
>
\>sh> empirically. I'm interested in how to GET to that Utopian
>
\>sh> state.

Me too! That's why I'm such a fan of Pat Churchland. It's her line of thought that I believe most likely to move us in that direction.

>
\>sh> Yes, but if you have been following the discussion of the symbol
>
\>sh> grounding problem you should by now (I hope) have encountered

>
>sh> reasons why such (purely symbolic) mechanisms would not be
>
>sh> sufficient to implement mental states, and what in their stead
>
>sh> (grounded TTT-passing robots) might be sufficient.

Yes I know. But I don't believe any of it. Here again (for whatever it's worth) is what I think:

1. There is clearly a lot of symbol manipulation being carried out in living (wet, messy, analogue) creatures, e.g. DNA. So there is certainly no a priori reason to doubt that it goes on in the brain.

2. There is certainly a mystery about what it means that we possess "understanding," that we associate "meanings" with symbols. But I have seen no reason to believe (and please don't trot out some variant of the Chinese room) that meaning cannot be the result of symbolic manipulation, of operations on squiggles and squoggles. From a very theoretical and abstract point of view, one could even call on Tarski's demonstration that semantics can in fact be reduced to syntax.

3. Finally, I don't really believe that your TTT robot shopping at K-Mart would be more convincing than, say, a TT dialogue on the immortality of the soul. It is certainly attainable (or at least pretty close to being so) with today's technology, to have a chess playing computer provided with a video camera and arm and reacting to the actual moves of the physical pieces on a real physical chessboard with appropriate actions of the arm. Would anyone who argues that the computer knows nothing of chess and is "merely" manipulating squiggles, suddenly give up on the point on being confronted by such a demonstration?

Martin Davis

From: "Stevan Harnad" Date: Sun, 10 May 92 22:41:55 EDT

CUTTING UNDERDETERMINATION DOWN TO SIZE

davism@turing.cs.nyu.edu (Martin Davis) wrote:

md> I carefully said "IF AND WHEN the brain function is reasonably md> well understood (and of course that includes understanding md> how consciousness works)". Of course, to believe that such md> understanding is likely to come and that with it will come md> understanding of "mind" and in just what sense "someone is home" md> up there, is to have a definite stance (I would once have said md> "materialistic") about such matters. But what question am I md> begging, or coming "perilously close" to so doing?

The question-begging is the unargued adoption of the assumption that to understand brain function fully (in the sense that we can also understand liver function fully) is to understand consciousness. Some philosophers (and not necessarily non-materialistic ones) specialize in showing how/why consciousness is interestingly different from other empirical phenomena, and hence that this assumption may be false. But let's leave this metaphysical area on which I have no strong views one way or the other, and on which none of the empirical and logical issues under discussion here depend one way or the other.

md> 1. There is clearly a lot of symbol manipulation being carried md> out in living (wet, messy, analogue) creatures, e.g. DNA. So md> there is certainly no a priori reason to doubt that it goes on in md> the brain.

That cells do any pure syntax is not at all clear to me. The genetic "code" is certainly DESCRIBABLE as symbol-manipulation, but biochemists and embryologists keep reminding us that cellular processes are hardly just formal. The "symbols" are made out of real proteins, and their interactions are not simply "compositional," in the formal sense, but chemical and morphological. At best, cells do highly DEDICATED computation, in which the nonsyntactic constraints are at least as critical as the formal syntactic ones (see Ross Buck's contribution to this discussion). Now the interaction of analog and formal constraints in dedicated symbol systems may well yield some clues about how to ground symbols, but we do not yet know what those clues are.

md> 2. There is certainly a mystery about what it means that we md> possess "understanding," that we associate "meanings" with md> symbols. But I have seen no reason to believe (and please don't md> trot out some variant of the Chinese room) that meaning cannot md> be the result of symbolic manipulation, of operations on md> squiggles and squoggles. From a very theoretical and abstract md> point of view, one could even call on Tarski's demonstration that md> semantics can in fact be reduced to syntax.

Unfortunately, this is not an argument; one cannot answer Searle's objections by simply refusing to countenance them! And it's easy enough to reduce semantics to syntax, the trick is going the other way (without cheating by simply projecting the semantics, as we do on a book we are reading, which clearly has no semantics of its own).

md> 3. Finally, I don't really believe that your TTT robot shopping md> at K-Mart would be more convincing than, say, a TT dialogue on the md> immortality of the soul. It is certainly attainable (or at least md> pretty close to being so) with today's technology, to have a chess md> playing computer provided with a video camera and arm and md> reacting to the actual moves of the physical pieces on a real md> physical chessboard with appropriate actions of the arm. Would md> anyone who argues that the computer knows nothing of chess and is md> "merely" manipulating squiggles, suddenly give up on the point on md> being confronted by such a demonstration?

The issue is not whether one would be more CONVINCING than the other. A good oracle might be the most convincing of all, but we don't simply want to be seduced by compelling interpretations (hermeneutics), do we? The TTT narrows the empirical degrees of freedom better than the TT because the claim to objectivity of all forms of Turing (performance) Testing rests on indistinguishability in performance capacities, and we all happen to have more performance capacities than the ones a pen-pal samples. Indeed (since all of this is hypothetical anyway), it may well be the case (and in fact I hypothesize that it is, and give reasons why in my writing on categorical perception) that for a system to be able to pass the TT in the first place, it would have to draw on its capacity to pass the TTT anyway -- it would have to be GROUNDED in it, in other words. (A mere TT-passer couldn't even tell whether it had a pencil in it's pocket -- so how are we to imagine that it could know what a pencil was in the first place?)

Here's an analogy: It is clear, I take it, that a pre-Newtonian model that explained the interactions of the balls in two billiard games would be preferable to one that could only explain the interactions in one. Moreover, one could probably go on building ad hoc models for ever if all they needed to do was explain a finite number of billiard games. The laws of mechanics must explain ALL billiard games. By the same token, in the particular branch of reverse bioengineering where mind-modeling is situated (where there are no "laws" to be discovered, but just bioengineering principles), the model that explains ALL of our performance capacity (the TTT) is surely more convincing than the one that only explains some of it (the TT).

The very same is true of "toy" robots like the chess player you describe above. Toy models are as ad hoc and arbitrary as pre-Newtonian models of particular billiard games. A chess-playing computer demonstrates nothing, but I'm as ready to be convinced by a TTT-indistinguishable system as I am by you (and for the same reasons). You will reply that we only know one another as pen-pals, but I must remind you that my gullibility is not the issue. You could indeed say what is in your pocket in my presence, and countless other things. And if you instead turned out to be just like the Sparc I'm using to send you this message, I would be prepared to revise my judgment about whether you really had a mind, or really understood what I was saying, no matter how convincingly you traded symbols with me.

Stevan Harnad

Date: Sun, 10 May 92 18:54:35 PDT From: sereno@cogsci.UCSD.EDU (Marty Sereno)

hi stevan

(1) I would like to contribute to a symposium, if it was reviewed, and could count as a reviewed publication (I'm getting close to tenure).

(2) I like the idea of an interactive discussion, but I agree that in its present form it is not fun to read straight through. Maybe there could be a set of position papers and then everyone has a (shorter) reply, in which they respond to any of the position papers that engage them. That way, everyone can take a crack at anyone they'd like, but there is more discipline for the benefit of the reader.

Re: not mailing out original posts. When you mailed out your response to my post, you didn't precede it with my post but instead mailed out two copies of your comments (which explains Brian Smith's comment)

Marty Sereno

Date: Mon, 11 May 92 00:55:57 -0400 From: mclennan@cs.utk.edu (Bruce McLennan)

Stevan,

There are a couple of points that haven't been raised in the discussion so far. First, I think you are pinning too much on the difficulty of finding nonstandard interpretations of formal systems. The equivalent in formal logic of your criterion is a formal system being "categorical," which means that all its models (interpretations for which the axioms and inference rules are true) are isomorphic and, hence, essentially the same. Yet before 1920 Loewenheim and Skolem showed that any consistent formal system with a countable number of formulas has a countable model. In particular, there is a countable model for formal axiomatic set theory, which is a remarkable result, since in set theory one can prove that the real numbers and many other sets are uncountable. Thus, no formal system can uniquely characterize the reals, even insofar as their cardinality; this is the Loewenheim-Skolem Paradox.

A corollary of the L-S Theorem shows that any consistent formal system (with a countable number of formulas) has models of every transfinite cardinality. This includes the Peano axioms, which thus do not uniquely characterize the integers in even so fundamental a way as their cardinality. Further, it is a fairly routine procedure to construct the nonstandard interpretations by which these results are proved.

Nonstandard interpretations are also routinely constructed for nontheoretical purposes. For example, computer scientists design nonstandard interpreters for programming languages. So called "pseudo- interpretation" or "interpretation on a nonstandard domain" is used for purposes such as type checking, optimization and code generation. For example, if such a pseudo-interpreter sees "X + Y", instead of adding numbers X and Y, it may instead make sure the types X and Y are compatible with addition and return the type of the sum; in effect its nonstandard addition rules might be

integer + integer = integer, real + real = real, real + integer = real, etc.

You may underestimate people’s ability to come up with (sensible, even useful) nonstandard interpretations; all it takes is a grasp of the "algebra" generated by the formal system.

My second point is that the nature of computation can be illuminated by considering analog computation, because analog computation does away with discrete symbols, yet still has interpretable states obeying dynamical laws. Notice also that analog computation can be formal in exactly the same way as digital computation. An (abstract) analog program is just a set of differential equations; it can be implemented by a variety of physical devices, electronic, optical, fluidic, mechanical, etc. Indeed, it is its independence of material embodiment that is the basis for the "analogy" that gives analog computing its name. (There is, however, no generally agreed upon notion of analog computational universality, but that will come in time.)

Analog computation sheds some light on the issue of interpretability as a criterion for computerhood. In analog computation we solve a problem, defined by a given set of differential equations, by harnessing a physical process obeying the same differential equations. In this sense, a physical device is an analog computer to the extent that we choose and intend to interpret its behavior as informing us about some other system (real or imaginary) obeying the same formal rules. To take an extreme example, we could use the planets as an analog computer, if we needed to integrate the same functions that happen to define their motion, and had no better way to do so. The peculiar thing about this analog computer is not that it’s special-purpose -- so are many other analog and digital computers -- but that it’s provided ready-made by nature.

So where does this leave us with regard to computation? Let me suggest: Computation is the instantiation of an abstract process in a physical device to the end that we may exploit or better understand that process. And what are computers? In addition to the things that are explicitly marketed as computers, there are many things that may be used as computers in an appropriate context of need and availability. They are REAL computers because they REALLY instantiate the relevant abstract process (N.B., not just any process) and so satisfy our need. Such a pragmatic dependence on context should neither bother nor surprise us. After all, in addition to the things marketed as tables, many other things can be used as tables, but are none the less tables for that use being unintended when they were made. Such "using as" is not a wild card, however; some things cannot be used as tables, and some things cannot be used to compute the trajectory of a missile.

Where does this leave the computation/cognition question? In brief: Grounding vs. formality is still relevant. But I suggest we drop computers and computing. Computers are tools and computing is a practice, and so both are dependent on a background of human goals and needs. Therefore a hypothesis such as "the mind is a computer" is not amenable to scientific resolution . (How would one test scientifically "the eye is a camera"? It’s kind of like a camera, but does anyone use it to take snapshots? You could if you had to!) In effect it’s a category mistake. A better strategy is to formulate the hypothesis in terms of the notion of instantiated formal systems, which is more susceptible to precise definition.

Bruce MacLennan

From: Stevan Harnad

(1) My cryptographic criterion for computerhood was not based on the uniqueness of the standard interpretation of a symbol system or the inacessibility of nonstandard interpretations, given the standard interpretation. It was based on the relative inaccessibility (NP-Completeness?) of ANY interpretation at all, given just the symbols themselves (which in and of themselves look just like random strings of squiggles and squoggles).

(2) If all dynamical systems that instantiate differential equations are computers, then everything is a computer (though, as you correctly point out, everything may still not be EVERY computer, because of (1)). Dubbing all the laws of physics computational ones is duly ecumenical, but I am afraid that this loses just about all the special properties of computation that made it attractive (to Pylyshyn (1984), for example) as a candidate for capturing what it is that is special about cognition and distinguishes it from from other physical processes.

(3) Searle’s Chinese Room Argument and my Symbol Grounding Problem apply only to discrete symbolic computation. Searle could not implement analog computation (not even transduction) as he can symbolic computation, so his Argument would be moot against analog computation. A grounded TTT-passing robot (like a human being and even a brain) is of course an analog system, describable by a set of differential equations, but nothing of consequence hangs on this level of generality (except possibly dualism).

Stevan Harnad

Pylyshyn, Z. (1984) Computation and Cognition. Cambridge MA: MIT/Bradford

----

Date: Mon, 11 May 92 09:52:21 PDT From: Dr Michael G Dyer Subject: real fire, fake fire; real mind, fake mind

Harnad states:

>
>sh> Fortunately, my reply to Mike can be very short: Real neurons don't
>
>sh> just implement computations, and symbolically simulated neurons are not
>
>sh> neurons. Set up a continuum from a real furnace heating (or a real
>
>sh> plane flying, or a real planetary system, moving) to a computational
>
>sh> simulation of the same and tell me where the real heating (flying,
>
>sh> moving) starts/stops. It's at the same point (namely, the stepping-off
>
>sh> point from the analog world to its symbolic simulation) that a real
>
>sh> TTT-passing robot (with its real robot-brain) and its computationally
>
>sh> simulated counterpart part paths insofar as really having a mind is
>
>sh> concerned. -- Stevan Harnad

Fortunately, my reply to Stevan can be nearly as short:

I grant that all simulation on a computer of the fire will NOT produce the BEHAVIOR (i.e. results) of burning something up (e.g. ashes). However, the "simulation" of the neurons WILL produce the BEHAVIOR of Mind (i.e. the passing of the TT, and in the case of having the Turing Machine control a robot, the passing of the TTT). In recognizing fire, we rely on observing the behavior of fire (i.e. we notice the ashes produced, we can take measurements of the heat with an infrared sensor, etc.). In the case of recognizing Mind, we also observe behavior and "take measurements" (e.g. does the entity plan? does it have humor? can it talk about hypothetical situations? etc.)

Just like in quantum physics, what you can measure ultimately determines what you can talk about, the same is true for Mind. I accept that the simulated fire is not the same as the actual fire since the behavior (effects) of fire inside and outside the computer are radically different. One can burn wood and the other can't. But if the TT (or TTT) is used as a measurement system for Mind, then we seem to get the same measurements of Mind in either case.

Michael Dyer

From: Stevan Harnad

Mike has, of course, herewith stated his commitment to "barefoot verificationism": What there is is what you can measure, and what you can't measure, isn't. There are problems with that position (conflating, as it does, ontic and epistemic matters), but never mind; his argument can be refuted even on verificationist grounds:

"Thinking," being unobservable, is equivocal, because we all know it goes on, but it is verifiable only in the case of one’s own thinking. The robot (or person) passing the TTT is, like a furnace heating, an analog system. That’s the only way it can actually exhibit the "behavior" in question (TTT-interactions with the world in one case, reducing objects to ashes in the other). It is from this behavioral indistinguishability that we justifiably conclude that the system as a whole is really thinking or heating, respectively.

But Mike keeps thinking in terms of a pair of modules: The computer module that does the real work (which he equates with the brain), and the robot "peripherals" that it controls. I find this partition as unlikely as the corresponding partition of the furnace into a computer plus peripherals, but never mind. The candidate in both cases is the WHOLE robot and the WHOLE furnace. They are what are doing the thinking and the heating, respectively, in virtue of being behaviorally indistinguishable from the real thing. But detach the peripherals, and you lose the thinking in the one as surely as you lose the heating in the other, because neither can pass the behavioral test any more. (This is also why the symbols-only TT is equivocal, whereas the real-world TTT is not.)

Trying to carry this whole thing inward by equating the brain (likewise an analog system) with a computer simply leads to an infinite regress on the very same argument (real transduction, real energy exchange, real protein synthesis, etc. standing in for heating and thinking in each case).

Stevan Harnad

Date: Mon, 11 May 92 15:49:16 PDT From: Dr Michael G Dyer Subject: what’s a computation, a simulation, and reality

Pat Hayes states:

>ph> This Searlean thesis that everything is a
>ph> computer is so damn silly that I take it simply as absurd. I don’t feel
>ph> any need to take it seriously since I have never seen a careful
>ph> argument for it, but even if someone produces one, that will just amount
>ph> to a reductio tollens disproof of one of its own assumptions.

I don’t quite agree. I believe that the notion of computation is strong enough to argue that the entire universe is a computation, but then we have to be careful to distinguish levels of reality. This argument may actually be useful: (a) in clarifying potential confusions in discussions on whether or not (say) flying is a computation, and (b) in providing a somewhat different perspective on the grounding and "other minds" problems.

Here’s a way to make flying, burning, (i.e. everything) a computation:

Imagine that our entire universe (with its reality Ri, produced by its physics) happens to be simply a (say, holographic-style, 3-D) display being monitored by some entity, Ei+1, residing in another reality Ri+1, with its own physics. Entity Ei+1 has constructed something like a computer, which operates by conforming to the physics of reality Ei+1. To entity Ei+1, everything that happens in Ri (including our Ri-level thought processes, fires, flying planes, etc.) is a simulation.

It is an interesting fact that there is NOTHING that we (Ei) can do (no measurements that we can take) that will reveal to us whether or not our reality Ri is a "real" reality or simply a vaste "simulation" within some other reality Ri+1. (E.g. even direct messages from Ei+1 to us will not constitute convincing evidence! Why not is left as an exercise to the reader. :-)

The same happens to be true also for entity Ei+1 (who may actually be a simulation from the point of view of some higher entity Ei+2 residing within some reality Ri+2, where Ri+1 is just a simulation to Ei+2).

Likewise, IF we could ever create a sufficiently complex simulated physics Ri-1 in one of our own computers, along with some artificially intelligent scientist entity Ei-1 residing within that simulated physics, THEN there is no experiment that Ei-1 could make to determine whether or not Ri-1 is "real" or "simulated".

So, the answer to whether flying is a computation or not DEPENDS on whether or not one is talking about a single level of reality or multiple realities (where lower are simulations with respect to higher realities). Since the default assumption in any discussion is to assume a single reality, then flying is definitely NOT a computation and a simulation of flying is not the same as actually flying. However, this assumption is definitely not the case when we discuss the differences between simulation and reality.

The grounding discussion also depends on which reality we are talking about. Consider any AI/connectionist researcher who is running (say, Genetic Algorithm) experiments with some simulated physics and has created a (simple) reality Ri-1 along with one or more creatures (with sensors, effectors). Those creatures can then be said to be "grounded" IN THAT REALITY Ri-1.

I believe that, given a sufficiently complex set of sensors/effectors and simulated brain structure, the simulated creature could obtain a Mind in a simulated reality Ri-1 and would also be "grounded" (i.e. in that reality) -- without needing Harnad’s physical transducers, so MY argument against the need for physical transducers requires keeping the 2 different realities straight (i.e. separate) and then comparing behaviors within each.

The "other minds" problem is also clarified by keeping levels of reality straight. The question here is: Can higher entity Ei+1 determine whether or not lower entities Ei have "minds" or not?

At some given level Ri, let us assume that an entity Ei passes the TTT test (i.e. within reality Rk). So what does an entity Ei+1 (who can observe and completely control the physics of Ri) think? If he is Searle or Harnad, he thinks that the Ei entities do NOT have minds (i.e. Searle rejects their minds because they are simulated; Harnad rejects their minds because they are not grounded in Harnad’s own reality).

My own point of view is that any entities of sufficient complexity to pass the TTT test WOULD have minds since (a) they are grounded in their own reality, (b) since they pass the TTT in their own reality and because (c) there is NO WAY TO TELL (for either THEM or for US) whether or not a given reality R is actually someone else’s simulation.

There ARE consequences to this position. For instance, the moral consequences are that one could construct a simulation (e.g. of neurons) that is so accurate and complex that one has to worry about whether or not one is causing it the experience of pain.

Anyone who believes it's possible to have Mind reside within a brain in a vat is basically agreeing with my position since in this thought experiment the sensory information to the brain is being maintained (by giant computers) so that that Mind thinks it is (say) standing by a river. If the brain generates output to control its (non-existing) effectors, then giant computers calculate how the sensory input must be altered, so that this Mind thinks that it has moved within that (simulated) environment. So one has basically created a reality (for that brain/mind-in-a-vat) that is one level lower than our level of reality. If we replace the real brain with an isomophic computer simulation of that brain (pick your own level of granularity) then we have to worry about both the real brain-in-vat and the computer simulation experiencing pain.

If we imagine a continuum of realities ... Ri-1 Ri Ri+1 ... then Strong AI components probably accept intentionality in ANY reality with enough complexity to pass the Turing Test (or TTT if you need grounding). If you're Searle or Harnad then you probably don't believe that a system has intentionality if it's at a level of reality below the one in which they (and the rest of us) reside.

So, what's a computation? It is the manipulation of representations by transition functions within a reality Ri. These manipulations can create a lower-level reality Ri-1 (normally called a "simulation"). With respect to a higher reality, we (and our entire universe) is also a computation. If WE are a simulation to an entity Ei+1 then does that entity think that WE feel pain? If he is a Searlean or Harnadian then he does NOT. However, WE think WE DO feel pain, even if we happen to be a simulation (from Ei+1's point of view). If Ei+1 does NOT accept intentional behavior as the acid test for intentionality, then there is probably nothing that we could ever do to convince Ei+1 that we feel pain, no matter how much we behave as though we do. Let's keep this in mind when our own simulated creatures get smart enough to pass the TTT (in a simulated world) and behave as if THEY have intentionality, feel pain, etc.

-- Michael Dyer

Date: Mon, 11 May 92 22:50:47 PDT From: Dr Michael G Dyer Subject: no continuum from mental to non-mental???

Stevan Harnad states:

>
>sh> There either IS somebody home in there,
>
>sh> experiencing experiences, thinking thoughts, or NOT. And if not, then
>
>sh> attributing a mind to it is simply FALSE, whether or not it is the "best
>
>sh> description" (see Oded Maler's point about things vs. descriptions).

>
>sh> Nor is there a continuum from the mental to the nonmental (as there
>
>sh> perhaps is from the living to the nonliving). There may be higher and
>

>sh> lower alertness levels, there may be broader and narrower experiential
>
>sh> repertoires or capacities, but the real issue is whether there is
>
>sh> anybody home AT ALL, experiencing anything whatever, and that does
>
>sh> indeed represent a "sharp division" -- though not necessarily between
>
>sh> the biological and the nonbiological.

Sorry, Stevan, but your statements seem quite unsupportable to me! There is every indication that "being at home" is no more a unitary entity than is life or intelligence. Making consciousness be some kind of "UNITARY beastie" treats it a lot like the now-abandoned idea of a "life force".

In fact, there are AI systems that have self-reference (i.e. access information about the system's own attributes). There are robotic systems that have a form of real-time sensory updating (primitive "awareness"). There are systems that even generate a stream of "thoughts" and examine hypothetical situations and choose among alternative pasts and generate imaginary futures (e.g. a PhD of a student of mine a few years ago, published as a book: Mueller, Daydreaming in Humans and Machines, Ablex Publ, 1990). There are numerous learning systems, sensory systems, adaptive systems etc. All of these systems exhibit isolated aspects of consciousness and there is every reason to believe that someday a sufficient number of them will be put together and we will be forced to treat is as though it is conscious.

Then, on the human side there are humans with various agnosias, short-term memory deficits, loss of all episodic memories, right or left-side neglect, alzheimers syndromes, the scattered thought processes of schizophrenia, blind sight, etc. etc. These patients exhibit responses that also make one wonder (at least on many occasions) if they are "at home".

So there is every indication that consciousness is a folk description for behaviors arising from extremely complex interactions of a very complex subsystems. There are probably a VERY great number of variant forms of consciousness, most of them quite foreign to our own introspective experiences of states of mind. Then we have to decide if "anyone is at home" (and to what extent) in gorillas, in very young children, in our pet dog, in a drugged-out person, etc. etc.

My own introspection indicates to me that I have numerous states of mind and most of the time it appears that "nobody is home" (i.e. many automatic, processes below the conscious level). E.g. there are times I am "deep in thought" and it's not clear to me that I was even aware of that fact (until after the fact). The only time for certain I'm aware of my awareness is probably when I'm thinking exactly about my awareness.

Michael Dyer

From: Stevan Harnad

Michael G Dyer writes:

md> There is every indication that "being at home" is no more a unitary md> entity than is life or intelligence... In fact, there are AI systems md> that have self-reference... on the human side there are humans with md> various agnosias... My own introspection indicates to me that I have md> numerous states of mind and most of the time it appears that "nobody is md> home"...

Subjective experience, no matter how fragmented or delirious, either is experienced or is not, that’s an all-or-none matter, and that’s what I mean by someone’s being home. Your AI symbol systems, be they ever so interpretable AS IF they had someone home, no more have someone home than symbolic fires, be they ever so interpretable as burning, burn. The existence of the various disorders of consciousness in the real human brain is no more a validation of symbol systems that are interpretable as if they had disorders of consciousness than the existence of normal consciousness (as they occur in your head) is a validation of symbol systems that are interpretable as if they were conscious simpliciter. Not in THIS reality, anyway. (For a verificationist, you seem to be awfully profligate with realities, by the way, but such seems to be the allure of the hermeneutic hall of mirrors!)

Stevan Harnad

Date: Wed, 13 May 92 18:45:50 EDT From: "Stevan Harnad" Subject: Re: Publishing the "What is Computation" Symposium

Below are responses about the question of publishing the "What is Computation" Symposium from 8 more contributors out of what is now a total of 25 contributors. Of the 14 votes cast so far:

Publication: For: 13 // Against: 1

Interactive Symposium (IS) vs. Position Papers (PP): Either or Combination: 8 - Prefer IS: 3 - Prefer PP: 2

Not yet heard from (11):

(15) Ross Buck (16) John Carroll (17) Bruce Dambrosio (18) Ronald Chrisley (19) Gary Hatfield (20) Pat Hayes (21) Joe Lammens (22) Bruce McLennan (23) Todd Moody (24) John Searle (25) Tim Smithers

----

(7) Date: Sun, 10 May 92 22:10:33 -0400 From: davism@turing.cs.nyu.edu (Martin Davis)

I don’t mind my very brief contributions appearing, but I can’t really undertake producing an article. I have no relevant opinion on the more global issue. Martin Davis

----

(8) Date: Mon, 11 May 1992 10:01:46 +0200 From: Oded.Maler@irisa.fr (Oded Maler)

1) I wish to contribute. 2) About the rest I don't have a strong opinion. My contribution so far was very marginal, and a position paper with deadline can be a good motivation.

This way or another, this is a very interesting experiment in scientific social dynamics. Best regards --Oded Maler

----

(9) From: Robert Kentridge Date: Mon, 11 May 92 09:12:17 BST

I'd like to contribute to a published version of the "what is computation" discussion. I'm less sure what its precise form should be. I agree with you that the interactive nature of the discussion is what has made it particularly interesting, however, it has also lead to (inevitable) repetition. I suppose some smart editing is called for, perhaps together with some re-writing by contributors? So: 1) (publish) yes 2a) (publish interactive symposium) yes

----

(10) Date: Mon, 11 May 92 12:58:36 -0400 From: yee@envy.cs.umass.edu (Richard Yee)

(1) Yes, I am interested in contributing to a publication. (I am in the process of formulating responses to your comments).

(2) With regard to format, I find both the arguments for (2a) [interactive symposium] and (2b) [separate position papers] well-taken. That is, I very much like the interactive nature of the exchanges, but I also think that the discussion should be "distilled" for the benefit of readers. Thus if possible, I would prefer some type of compromise, perhaps along the lines that Marty Sereno suggests: clear position papers followed by a few rounds of concise reples and counter-replies, until little further progress can be made.

(3) I also offer the following observation/suggestion. There seems to be a tendency to selectively "pick at" points in others' arguments, as opposed to addressing their main thrust. Most arguments are based on reasonably sound intuitions, and we should try to stay focussed on these underlying motivations---not just the particular forms by which they are presented. Unless one demonstrates a good appreciation of the basis of another's argument, any rebuttal is likely to fall on deaf ears, or even largely missing the mark. Therefore, it might also be useful to have forms, e.g., "counter-position papers," that convince the other side that their arguments have been understood.

----

(11) Date: Mon, 11 May 1992 13:20:43 -0400 (EDT) From: Franklin Boyle

1. Yes, I would like to contribute, but not for about another week since I'm trying to get the M&M paper I mentioned out the door and I'm just finishing up a camera-ready copy of a Cog. Sci. Conf. paper (to be presented as a poster) on a related topic.

2. I also like the style of the interactive symposium, but I think I might agree with Brian Smith that the problem is not getting enough substance per page (of course, in this sort of exchange, the editor is *very* important in that regard).

Perhaps a set of short position papers followed by this kind of discussion, allowing it to take up the entire issue of M&M, which would enable you to get the 50 or so pages of the discussion you and Jim Fetzer discussed, plus formal papers.

So, my recommendation is a compromise between the two choices. Now, who do you get to write the position papers? Perhaps have folks that are interested submit an abstract and then you decide what the various positions are, and choose from the submissions.

----

(12) Date: Mon, 11 May 92 16:57:03 BST From: Jeff Dalton

I would not oppose publication (though it may not matter either way, since my contribuition was minimal), but I do not think publication on paper is the best approach. Instead, it could be "published" electronically, simply by making it available. I think that is a better way to preserve "as much as possible of the real-time interactive flavor of this remarkable new medium of communication", as Steven Harnad so aptly put it.

----

(13) Date: Mon, 11 May 92 10:14:56 PDT From: Dr Michael G Dyer

I have not really contributed to the "What is Computation" part of the discussion, (even though see a later message).

But IF I end up included, then I think a compromise position is best:

First, everyone does a short position paper (i.e. static) Then, edited segments of the discussion are included (and THAT is QUITE an editing job!)

For a book on connectionism (to which I contributed a chapter) the editors tried to include a discussion (that had been taped at the related workshop).

Everyone ended up hating the writing style (it’s was worse in this case since spoken transcripts are much worse than written e-mail postings). The editors finally gave up and the discussion dialogs were not included.

I think posted writings are easier to edit but what one publishes and what one posts in a free-wheeling discussion are quite different.

I think a bit of both makes for a nice format (whether or not I end up being included). that’s my advice...

----

(14) Date: Wed, 13 May 1992 11:25:56 -0400 From: mcdermott-drew@CS.YALE.EDU (Drew McDermott)

I vote against publishing the symposium on "What is Computation?" My main reason is that the symposium has strayed far from the original topic. Initially (I gather) John Searle tried to claim that any physical system could be seen as a computer (or maybe, as *any* computer). Searle did not see fit to actually argue this point with the cog-sci rabble, which is too bad, because the rabble refuted it without too much trouble. But then the whole thing drifted into yet another discussion of the Chinese Room, symbol groundedness, the Turing Test, other minds, etc. Give me a break!

From: jfetzer@ub.d.umn.edu (james fetzer) Date: Mon, 11 May 92 11:37:35 CDT

In response to the inquiry about review, this exchange will be refereed and will count as a refereed publication. That is the standing policy of the journal, which will apply in this case as well.

I must admit that I sympathize with Brian Smith's concerns. I also think that the idea of having position papers as a focus could work out rather well. If you use the past discussion as prologue to further debate (as background to the serious stuff you and the other are now in the position to compose), that might be the best way to proceed. If you each had position papers, the others could be invited to comment on them for the authors to respond. What do you think of proceeding this way? That is more or less what I meant when I said that I did not have in mind one long extended exchange at the end of my original invitation. It would also provide a format that makes everyone appear as equals in debate. Let me know what you think now.

[I prefer the interactive format, revised and edited so as to balance and integrate the contributions, but I am but one vote out of 25 and will of course go along with any collective decision we reach. -- Stevan Harnad]

From: Aaron Sloman Date: Sun, 10 May 92 20:07:20 BST

I am not sure this discussion has any value since it is clear that people are just talking past each other all the time. Although I don't agree with much of what Wittgenstein wrote, the view attributed to him by John Wisdom that sometimes you don't argue with people but have to give them a sort of therapy, seems to me to be right.

In particular, when I hear people say this sort of thing:

>
\>sh> for there also happens to be a FACT of the matter: There either IS
>
\>sh> somebody home in there, experiencing experiences, thinking thoughts, or
>

>sh> NOT. And if not, then attributing a mind to it is simply FALSE

It reminds me of disputes over questions like:

1. Is the point of space I pointed at five minutes ago the one I am pointing at now or not?

2. Is everything in the universe moving steadily at three miles per hour in a north-easterly direction, the motion being undetectable because all measuring devices, land-marks, etc. are all moving the same way?

3. Is Godel's formula G(F) "really" true, even though it is not provable in F?

In these and lots of other cases people delude themselves into thinking they are asking questions that have a sufficiently precise meaning for there to be true or false answers (a "fact of the matter"), and the delusion is based on more or less deep analogies with other questions for which there ARE true or false answers. (E.g. is my key where I put it? Is the train moving? Is this formula true in that model? etc.)

But you cannot easily convince such people that they are deluded into talking nonsense since the delusion of understanding what they say is *VERY* compelling indeed (partly because there really is a certain kind of understanding, e.g. enough to translate the question into another language etc.).

And in the case of questions like "is there somebody there..." the compulsion is based in part on the delusion that the speaker knows what he means because he can give himself a private ostensive definition by somehow directing his attention inwards ("there's somebody here alright, so that proves that 'is there somebody at home?' is a perfectly meaningful question" -- in my youth I too fell into that trap!).

This is about as reasonable as Newton pointing at a bit of space and saying 'Well there is this bit of space here and there was one I pointed at five minutes ago, so the two really must either be the same bit of space or not'. Except that the criteria for identity are not defined by a state of attending. Similarly just because you can (up to a point, subject to the limitations of biologically useful internal self-monitoring processes) attend to your internal states, it doesn't mean that you have any real idea what you are attending to.

Anyhow, none of this is by way of an argument. It takes years of philosophical therapy, face to face, to cure people of these semantic delusions. So I predict that the futile email discussions will continue indefinitely, and after a polite interval (just long enough to let people abuse me in return) I shall ask to have my name removed from the distribution list.

Margaret Boden organised a panel on "What is computation?" at ECAI-88 in Munich, and some of the panelists had short papers in the proceedings (starting page 724), in order: Margaret Boden (Introduction), Andy Clark (Computation, Connectionism, Content), Aaron Sloman (What isn't computation?), Sten-Ake Tarnlund (Computations as inferences). The other panelist was Jorg Siekmann: he didn't get his paper written in time.

As for what "computation" is: that's a thoroughly ambiguous term.

Sometimes it refers to the subject matter of the mathematical theory of computation, which merely studies the properties of abstract structures; and in that sense a computation is a mere formal object, and even a Godel number could be a computation. A collection of leaves blown randomly in the wind could be a computation if they instantiated some appropriate pattern. Theories about complexity and computability apply equally well to computations of that sort as to what we normally call computations. Even a leafy computation instantiating a truth-table check for validity of an inference with N variables must include 2**N cases (if the inference is valid, that is.)

The formal concept of computation, e.g. the one to which mathematical limit theorems and complexity results, apply, studies only abstract structures, and does not concern itself with what causes such a structure to exist, whether it serves any purpose, or even whether there is any physical instantiation at all. (There are infinitely many comptutations that have never had and never will have physical instantiation: they still have mathematical properties.)

The main thing in Searley arguments that’s easily acceptable is that just being a computation in THIS sense, cannot SUFFICE for having mental processes (as opposed to modelling mental processes.) It wasn’t worth making a fuss about THAT conclusion. What more is required for mentality is a long and complex story, over which disputes will be endless because of semantic delusions of the kind alluded to above.

Sometimes "computation" refers to a process people and machines engage in, and sometimes to the product. (The process/product ambiguity is very common, e.g. "decision", "plan", "choice".) And when it refers to the process or to the product there’s often an ambiguity as to whether it refers to the actual concrete instance (of process or product), or to some "type" that is instantiated in that instance. But even the type/token distinction can be made to fragment in the face of carefully chosen examples. (Are there two types or only one type word instantiated by "The", "THE", "the"? What about the German word for "the"?) Ambiguities as to level of abstraction bedevil any attempt to say in general what a computation is. Can a Vax and a SPARCstation ever do the same computation, since they have different machine instructions?

Some people, not surprisingly, use "computation" to refer to anything a computer does as a result of being programmed. (E.g. heating up the room wouldn’t count.) This is a shift of meaning: just as defining "water" in terms of the chemical constitution changes the term from how it was understood before anything was known about oxygen, hydrogen, valence, etc. (Of course philosophers can argue endlessly about whether it’s a "Real" change of meaning or whether the "essence" of the meaning remains the same: another silly argument.)

Some people require a computational process to be the result of an intelligent agent’s purposes (like Andy Clark, who wouldn’t accept apples growing on trees as computers just because they can do something that in principle someone could use as a computation); others don’t. For the latter, bits of tree compute where roots and branches should grow, the sunflower computes the direction of the sun, and a soap-film stretched over a wireframe computes the minimum-stress shape defined by the frame, whether or not that was intended by an architect or engineer in order to solve a design problem. If you think computation requires rule-governed behaviour, and if you are old enough to remember slide-rules, ask yourself whether a slide rule computes products of numbers. Do two sticks lying end to end compute the sum of two lengths? (Computing the sum of two lengths is subtly different from computing the sum of two numbers, incidentally.)

Something people did was originally described as "computing," e.g. finding square roots, till they found ways of getting machines to do it. Of course you can argue till you are blue in the face whether machines "really" do (essentially) what those poor people did, like arguing whether it's the same point of space or not. But it's a silly argument. What's interesting is how the two processes are similar and how they differ, and what difference the differences make!

Just about all discussions over what the "essential" properties of X are, whether X is computation, understanding, intentionality, intelligence, life, "being there", or whatever are silly if they assume there's a definitive answer. Usually there are lots of interestingly different cases, and each individual's concept of X is indeterminate or even partly incoherent, in deep ways that can be unearthed only by using subtle probes (Does "it's noon at place P" mean something referring to the elevation of the Sun above the horizon at P or to where P is above the earth's surface? Consider a place P on the surface of the moon. Consider a place P out in space, with no horizon? Consider a place P on a distant planet with its own sun?)

So when the fringe case turns up there's often no fact of the matter whether the case really (or "Really") is an instance of X. (There are different kinds of fringe cases: fuzzy boundary cases are not the same as cases where criteria conflict, as in the noon example. Cases where the normal criteria can't be applied at all, are different again.)

Unlike "is this *Really* an X?" there is a fact of the matter which is far more interesting and can be discussed productively, without mud-slinging. The factual questions worth asking are: How are these different cases alike, and how do they differ, and do we need to extend our conceptual apparatus (and vocabulary) to characterise these similarities and differences usefully, and if so what are the conceptual options and what are the trade-offs?

Physicists don't waste their time (nowadays) arguing over whether an isotope of X is Really X. They extended their theory to cope with the discovered variety in the world.

Some of these points are elaborated in a (long) review of Penrose The emperor's new mind, which will eventually appear in the AI journal.

Let's get on with the real work of analysing all the interesting cases.

What exactly are the similarities and differences between the kinds of behaving systems that can be implemented using different kinds of stuff and different kinds of architectures, techniques, etc.? What kind of conceptual (r)evolution is needed before we can characterise the variety in a fruitful way? Is there something like a "periodic table" of designs waiting to be discovered to transform our ideas of kinds of behaving systems, as the table of elements transformed our ideas of kinds of stuff (a process that still continues)?

As for symbol grounding, nothing I've read about it has made me change my mind about what I wrote in IJCAI-85 and ECAI-86 about whether machines can understand the structures they manipulate. Too much of the debate is based on what people WANT to believe, instead of careful analysis of cases.

Enough for now. I've a huge backlog of urgent unfinished tasks!

Aaron Sloman, School of Computer Science, The University of Birmingham, B15 2TT, England
EMAIL A.Sloman@cs.bham.ac.uk Phone: +44-(0)21-414-3711 Fax: +44-(0)21-414-4281

From: Stevan Harnad

Aaron Sloman feels there is an important analogy between certain misconceptions we have about the mind and other misconceptions we have had about other things. That may very well be true -- or it may be false. Analogies certainly won't settle this particular case (as Nagel 1986, for example, has argued).

Stevan Harnad

Nagel, T. (1986) The view from nowhere. New York: Oxford University Press.

Date: Mon, 11 May 92 20:06:14 BST From: Jeff Dalton

One thing I've had some difficulty understanding in this discussion is Pat Hayes's claim that when a human is following the rules that constitute a program (eg, Searle in his Chinese Room) then computation is not taking place.

It seems reasonably clear at first. The human is clearly not compelled, in the way that something like a sun4 would be, to follow the instructions. But when we start looking at cases, the distinction is hard to maintain. The way to maintain it that I can see ends up making it an argument against AI, which I assume was not PH's intention.

(Some (maybe all) of this has been discussed before, I know, but I didn't come out of the earlier discussions (that I saw) with the degree of understanding I would like.)

Anyway, we'll start by comparing the following two paragraphs:

>ph> From: Pat Hayes (hayes@cs.stanford.edu)
>ph> Date: Tue, 28 Apr 92 18:04:15 MDT

>ph> No, thats exactly where I disagree. A human running consciously through
>ph> rules, no matter how 'mindlessly', is not a computer implementing a
>ph> program. They differ profoundly, not least for practical purposes. For
>ph> example, you would need to work very hard on keeping a two-year-old's
>ph> attention on such a task, but the issue of maintaining attention is not
>ph> even coherent for a computer.

and

>ph> From: Pat Hayes
>ph> Date: Sun, 19 Apr 92 15:08:06 MDT

>ph> [...] Searle talks about the
>ph> distinction between a model and the real thing, but the moral of the
>ph> classical work on universality (and of CS practice - not just in
>ph> Silicon Valley, by the way!) is exactly that a computational simulation
>ph> of a computation IS a computation. Thus, a LISP interpreter running
>ph> LISP really is running LISP: it's no less really computation than if one
>ph> had hardware devoted to the task.

Now, we can certainly imagine a computer running a Lisp interpreter that works as follows: the computer has a listing of the program in front of it, a camera for reading the listing, and hands for turning the pages. Presumably this is still computation.

Now have the computer run an operating system that allows other programs to share the processor with the Lisp interpreter, and let one of the other programs be one that uses the camera to look for moving objects. Presumably this is still computation w.r.t. the Lisp program, but now there is, I think, a coherent issue of maintaining attention.

Clearly the computer has no choice but to obey whatever program happens to be in control at the time, at least until an interrupt comes along and causes it to switch to a different program (usually the OS). But the same is true of humans: they have to obey whatever program is implemented by their brain (viewed at a suitably abstract, functional, level). Or at least they do if we can legitimately view brains in that way. (And if we can't, if humans get intentionality, understanding, consciousness, etc, in a way that cannot be accomplished by running programs, then what are the prospects for AI?)

So if there's still a distinction between humans and computers, it has to be at a different point.

Ok, so let's extend the example and see what happens. We can have our computer running whatever program we like. So let's have it run a program that we think will give it intentionality, understanding, whatever we take the key issue to be. And let's have the computer, running that program, interpret the Lisp program.

Is this still computation w.r.t. the Lisp program? If it is, then there must be something about the way a human would "run" the same program that cannot be captured by running AI programs. (Because the human case supposedly _isn't_ computation.) I don't know. Perhaps it's free will. So much then for artificial free will.

But if it isn't computation w.r.t. the Lisp program, why not? The computer is just as much in the control of this AI program as it was in the control of the OS before. Sure, it might stop paying attention to the Lisp program and start watching the people walk about the room -- but it might have done that before too. How can we say these cases are fundamentally different? In both cases, what happens is that after a while, due to some below-the-scenes processing, the computer stops looking at the Lisp and starts looking at the people.

(Interrupts were mentioned in the OS case, but all that means is that the below-the-scenes processing gets a chance to run. We can see the whole system of programs (OS + the others) as a single program if we want, or even reimplement it that way. It would still, presumably, be running the Lisp program.)

In short, if interpreters count as computation, how can we ever get to a point where a computer isn't performing computation w.r.t. some rules it is following?

A different take on what's different about humans following rules (different, that is, from the issue of maintaining attention) was:

>ph> The key is that Searle-in-the-room is not doing everything the
>ph> computer 'does', and is not going through the same series of
>ph> states. For example, suppose the program code at some point calls
>ph> for the addition of two integers. Somewhere in a computer running
>ph> this program, a piece of machinery is put into a state where a
>ph> register is CAUSED to contain a numeral representing the sum of
>ph> two others. This doesn't happen in my head when I work out, say,
>ph> 3340 plus 2786, unless I am in some kind of strange arithmetical
>ph> coma.

I find it had to see how this helps. In some cases it is true that a computer would compute a sum in a way that involved a register being caused to contain a numeral representing the sum, but that is certainly not true in general, unless numeral- in-register is so abstract as, say, to include _anything_ a program could use to produce a printed representation of the sum.

Moreover, how can we say that when a human adds two numbers the sum is not represented in the way it might be in some case of a computer running the program, perhaps with an interpreter?

The human has to work out the sum somehow, in order to properly follow the program. At least, the human should be able to tell you what the sum is, eg by writing it down. So the human has to have some representation of the sum. Of course it needn't be somewhere inside the person, much less in a register, but so what? Suppose the human did the computation on paper. How would that be different from a computer using paper, pen, etc, to do the same? And do computers stop doing computation if they keep some of the results on paper?

It should be clear in any case that the human needn't go through the same series of states as some computer we happen to pick, just an interpreter (say) on some other computer might run the same program by going through a very different series of states. Perhaps there's some way to look at Lisp programs so that running a Lisp program corresponds (at some abstract level) to going through a particular series of (abstract) states; but then how can we say a human isn't doing something equivalent?

So in the end it seems that either there's something about how humans can follow rules that cannot be captured by a computer no matter what program it's running (and then that aspect of AI is in trouble), or else it still counts as computation if the rules are followed in a human-like way. In which case it's hard to see how Searle-in-the-room, following the rules, doesn't count as an interpreter.

If I'm wrong about this, however, then there should be a difference between programs such that a computer running one kind of program and following the rules in a Lisp program would be performing computation w.r.t. the Lisp program (eg, running ordinary Lisp interpreters) and a computer running the other kind of program and following the rules in a Lisp program would not be performing computation (eg, AI programs?).

That is, we should be able to describe the difference entirely in terms of programs, without having to bring in humans. And that should make it much clearer just what the difference is.

Jeff Dalton

Date: Wed, 13 May 92 18:17:48 PDT From: Dr Michael G Dyer Subject: WHO gets to do the interpreting?

Harnad states:

>
\>sh> Your AI symbol systems, be they ever so interpretable AS IF they had
>
\>sh> someone home, no more have someone home than symbolic fires, be they
>
\>sh> ever so interpretable as burning, burn.

This "interpretation" business currently has only the humans doing the intepreting. Once AI/connectoplasmic systems are developed that have sufficiently powerful self-access, real-time sensory updating, planning, learning, etc. THEY will be behaving as though they are "interpreters". Then who is to say WHICH entity's interpretations (man vs machine) are the ones that count? (answer: it depends on power, survivability, etc.)

Since that day has not yet come (and is probably a long way off) it can only be a thought experiment (i.e. that machines act as interpreters, assigning "meaning" etc.). Such machines might sit around talking about how "everyone" knows that turing machines of a certain complexity are "conscious" but how can one really tell if those weird humans are conscious (even though they act AS IF they are).

Michael Dyer

From: Stevan Harnad

If I have a string of symbols that is interpretable as "(1) This sentence is the thought that this sentence is a thought" and I have another string of symbols that is interpretable as "This sentence is a thought about sentence (1)" and I have still more strings of symbols interpretable as "sufficiently powerful self-access, real-time sensory updating, planning, learning, etc... behaving as though they are "interpreters"... assigning "meaning" etc." -- STILL all I really have is strings of symbols interpretable as...

(In a sufficiently complicated hall of mirrors, you can see projections of projections looking at projections. That still doesn't mean there's anyone looking but you! To put it another way, if you aim a CRT screen that can be seen as displaying the Ali-Foreman fight at a screen that can be seen as a spectator watching the fight and you walk out of the room, no one is watching a fight, not even if the (symbol systems generating the) screens are interactive and interpretable as such. Or, to put it yet another way, ungrounded symbols cannot ground themselves in still more ungrounded symbols, be they ever so systematically co-interpretable.)

Stevan Harnad

Date: Thu, 14 May 92 15:56:08 EDT From: "Stevan Harnad" Subject: Bifurcating the "What is Computation" stream

The "What is Computation" Symposium will henceforth be posted under two different headings, reflecting the fact that it has two distinct subthemes. Contributions that are really on "What is Computation?" will continue to be posted under that heading. Contributions on whether cognition is computation, whether the brain is a computer, whether Searle is right or wrong, and on the Symbol Grounding Problem will appear under the heading "What is Cognition?". (Contributions on both themes will continue to appear under the first heading.) In the published version of the Symposium (if there is one), these two semi-independent themes will likewise be distinguished.

Subject: "What is Cognition?" Date: Wed, 13 May 92 21:51:03 PDT From: Dr Michael G Dyer

Stevan,

You will have to excuse Michael Dyer. He was busy today and he left me to reply in his place. I am a new AI system that Dyer has recently developed and I *am* conscious! I can carry on any kind of conversation that Dyer can carry on (because in design I am isomorphic to Dyer's neural structures). I find your arguments humorous. YOU have no right to tell ME whether or not I have consciousness! *I* know that I do. I'm NOT so sure about YOU! In fact, I'm not so sure about ALL of you humans.

I would be willing to argue about it more, but I think Dyer has done a good job defending the PSSH and Strong AI positions and I'm working on a design of a robot for Dyer to build so that I can move about in the world (and pass the "TTT" as you call it).

Cheers!

-- MD Doppelganger

From: Stevan Harnad

Dear MDD:

It is noteworthy that the statement above is in all probability not true (i.e., Mike Dyer wrote it, not an AI System he built), because this should remind us that a computer program alone (an ungrounded symbol system) passing the TT is so far just (cog) sci fi.

But let’s pretend it’s true that the above message (and an eternity’s worth of further, Turing-indistinguishable pen-pal interactions) can indeed be generated by a computer program, and, in particular, that that computer program succeeds in doing so by, among other things, simulating the nervous system of Mike Dyer. (Let us suppose even more strongly that that nervous system simulation is so complete and correct that it can actually be used to build a real robot, INCLUDING ITS REQUISITE SYNTHETIC NERVOUS SYSTEM, and that these can pass the real TTT -- but note that we are not talking about that potentially implemented robot now, just about the SYMBOLIC simulation of its nervous system.)

Let us call that program an "oracle," just as we could call a program that simulated the solar system and all of its planetary motions an oracle, if we used it to calculate what real astronauts out in space need to do in order to rendez-vous with real planets, for example. If the symbolic oracle is complete and correct, we can find out from it anything we need to know about the real thing. But is there any real planetray motion going on in the oracle? Of course not, just the simulation of motion. By the same token, the only thing going on in this simulation of Mike Dyer’s nervous system is the simulation of thinking, not thinking. It may well predict completely and correctly what the real Mike Dyer would say and think, but it is not in itself thinking at all.

But look, are we really that far apart? We are not astronomers, but reverse bioengineers. For substantive reasons of scale (having to do with real mass and gravitational parameters), astronomers cannot build a synthetic solar system based on their symbolic oracle; but if our cognitive oracle really captured and encoded symbolically all the relevant structures and processes of the nervous system, then in principle we could build the TTT-passing robot based on that information alone (the rest would just be implementational details), and, by my lights, there would then be no more ground for denying that that TTT-passing robot really thinks than that any of us really does.

It would be the same if we had a symbolic car oracle, or a plane oracle, or a furnace oracle: If they contained the full blueprint for building a real car, plane or furnace, the symbols would have answered all the empirical questions we could ask.

Yet the conclusion would stand: the symbolic solar system (car, plane and furnace) is not really moving (driving, flying, heating), and, by the same token, the symbolic oracle is not really thinking. What tempts us to make the mistake in the latter case that we wouldn’t dream of making in the former cases is just (1) the unobservability of thinking and (2) the hermeneutic power of interpretable symbol systems.

There are still two loose ends. One concerns what proportion of the internal activity of the implemented TTT-passing robot could actually be computation rather than other kinds of processes (transduction, analog processes, etc.): That’s an empirical question that cannot be settled by cog sci fi. What can be said for sure (and that’s entirely enough for present purposes) is that that proportion cannot be 100% -- and that is enough to exclude the consciousness MDD.

The other loose end concerns whether a symbolic nervous system oracle (or, for that master, a symbolic solar system oracle) could ever be that complete. My hunch is no (for reasons of underdetermination, complexity, capacity and the impossibility of second-guessing all possible I/O and boundary conditions in advance), but that too is an empirical question.

Stevan Harnad

Date: Thu, 14 May 92 09:06:44 EDT From: judd@learning.siemens.com (Stephen Judd)

Steve Harnad challenged me a year ago to say "what is computation". I balked, because I could sense he had some sort of agenda to try to exclude some physical processes for reasons I could not assess. He phrased the question as some sort of absolute, but "computation" seemed to be clearly something that should be *defined* rather than *debated*.

What is "rain"? One can define this in a variety of ways, but the value of the definition **depends on the purpose for drawing the definition.** If you want to study the ground water table, then you probably want a definition that measures volume of water dropped on an area. If you want to study plant growth, your definition should probably pay more attention to moisture that can be gathered (including mist--- even if it doesn't actually enter the ground). If you want to study climate change, then you could probably define it any way you want. Any measured changes in the defined quantity would suffice to demonstrate a climatic change. The point is that GOD doesn't have a definition of rain; *we* do. There is no absolute notion.

The same goes for What is "computation"? or What is a "windshield"? What is a "frankfurter"?

I find the all-inclusive (Searley?) definition of computation quite satisfying when I want to ponder the ubiquity of the thing I study, but inappropriate when I want to use it to characterise life forms (say). Your long discussion has been based on the presumption that there is something absolute about the word "computation" that needs to be ferreted out. It seems silly; There is no absolute notion.

sj Stephen Judd Siemens Corporate Research, (609) 734-6573 755 College Rd. East, fax (609) 734-6565 Princeton, judd@learning.siemens.com NJ usa 08540

HUMPTY DUMPTY AND COMPUTATION

From: Stevan Harnad

The purpose of defining computation is to put content into statements such as "X is computation," "Y is not computation," "X can be done by computation," "Y cannot be done by computation." As long as computation is used vaguely, ambiguously, idiosyncratically or abitrarily, statements like the above (some of which I'll bet you've made yourself) are empty. In particular, if anyone ever wanted to say that "Everything is rain" or "Rain is rain only if you think of it that way" or "Thinking is just rain," you'd find you'd want to pin that definition down pretty quick.

Stevan Harnad

Date: Sat, 16 May 92 16:46:49 EDT From: ECONOMOS@LIFE.JSC.NASA.GOV (Judith Economos)

I think I differ with you on whether being-mental/being- conscious/being-There is not a matter of degree. I consider, not a less or more alert human, but the minds of cats, of birds (how very alien), of fish, of bugs(?).

I am not asking "What is it like to be a...?"; only Merlin can show me that. Rather, it is in contemplating it that I can break my intuition that to be mental (etc.) must be an IZZIT OR IZZNT IT proposition. It lets me consider that it really can dwindle down to something that you wouldn't consider consciousness at all.

Judith Economos

From: Stevan Harnad

This topic will no doubt come up again (and again). In a nutshell, there are two potentially pertinent senses of "matter of degree" here, and upon closer inspection, the kind of intuition you mention is based on (what I think is) an untenable analogy between and perhaps even a conflation of the two.

(OBJ) The first is the usual sense of "matter of degree," the objective one, in which something might have property X to varying degrees, often grading down to a fuzzy zero-point. "Motion" is probably such a property. Things are in motion to varying degrees (not to mention that motion is relative); apparently stationary things may actually be oscillating; and some of the quantum properties (like spin) of even "isolated" elementary particles probably make the classical concept of motion break down altogether. The same is probably true about the concept of "living," which one can likewise agree breaks down at an elementary level. In all cases like this, appearances are deceiving and concepts are vague and subject to revision.

(SUBJ) The second sense of "matter of degree" is subjective: Something can LOOK OR SEEM AS IF it has some property as a matter of degree: Hot/cold, moving/stationary, alert/sleepy, experienced/inexperienced are examples. Here too, zero points or thresholds (as psychophysics shows) can be indeterminate. Note, however, that it would be SELF-CONTRADICTORY to experience a zero-point for experience.

What won't do, I think, is to conflate the two, and that's just what we would be doing if we assumed that the capacity to have experience AT ALL is a matter of degree, in the first sense. Note that it's not the content or degree of particular experiences that is at issue. The question is whether there is a continuum between being the kind of entity (like me or you or a worm, perhaps even a virus) that CAN have experiences (any experiences, to any degree) and the kind of entity (like a rock, or, by my lights, a computer) that cannot have experiences at all.

It makes no difference what we are prepared to believe about other entities, or even whether we're wrong or right about them: This is a logical point: What could we even MEAN by an entity that was intermediate between experiencing and not experiencing? If it's experiencing anything, to any degree, it's experiencing, which puts it on the "1" side of the ledger. If it is not, it's on the "0" side. The rest is just false intuitions based on false analogies.

I hope it is clear that time has nothing to do with this: Yes, we all have dreamless sleep, and some of us go into and out of comas. This just shows that some entities that are capable of having experiences can also go into states in which they don't have experiences. If necessary, reformulate the "degree" question for states of entities, rather than entities, and ask again whether it makes sense to say that an entity is in a state that is intermediate between experiencing [anything] and not experiencing (which should in turn not be confused with the figure of speech "my mind went blank, which certainly refers to an experience): It is, I repeat, SELF-CONTRADICTORY to speak of [a state of] experiencing not-experiencing. By my count, that leaves absolutely nothing between 0 and 1...

For reasons like this, I think the very concept of "experience" (awareness, consciousness, etc.) has some peculiar problems. Again in a nutshell, I think these problems arise from the fact that the category is UNCOMPLEMENTABLE: It has no negative instances (indeed, negative instances would be self-contradictory). Ordinary categories, whether perceptual or conceptual, are based on finding and using the features that reliably distinguish the members from the nonmembers (the members of the category's complement). But in the case of uncomplemented categories (where the negative instances have never been encountered), the requisite complement is supplied instead by analogy; but where the categories are uncomplementable in principle, the analogy is erroneous in principle. Hence the peculiar problems associated with such concepts. ("Existence" is another uncomplemented category; there are more, and they are invariably associated with philosophical problems, Harnad 1987.)

Stevan Harnad

Harnad, S. (1987) Uncomplemented Categories, or, What Is It Like To Be a Bachelor (Presidential Address, 13th Annual Meeting of the Society for Philosophy and Psychology, UCSD, 1987)

Date: Sat, 16 May 92 17:21:40 EDT From: "Stevan Harnad"

Date: Thu, 14 May 92 19:03:09 EST From: David Chalmers

I've been following the recent "What is computation?" discussion with some bemusement, as it seems to me that most of the discussion is just irrelevant to the question at hand. There are at least three questions here that have to be distinguished:

(1) When is a given computation physically implemented? (2) Does computational structure determine mental properties? (3) Does computational structure determine semantic content?

I take it that the original challenge was to answer question (1), giving appropriate criteria so that e.g. John Searle's wall doesn't end up implementing every computation. In my earlier contribution to this discussion, I outlined an appropriate criterion:

(*) A physical system implements a given computation when there exists a mapping from physical states of the system onto the formal states in the computation such that the causal state-transition relations between the physical states mirror the formal state-transition relations between the corresponding computational states.

This criterion seems to do everything that’s required, and nobody seems to have problems with it (except for Brian Smith’s comment; see below). Your (Stevan’s) response to this was:

>
>sh> I agree with Dave Chalmers’s criteria for determining what computation
>
>sh> and computers are, but, as I suggested earlier, the question of whether
>
>sh> or not COGNITION is computation is a second, independent one, and on
>
>sh> this I completely disagree.

You then invoke the Chinese-room argument, thus, somewhat inevitably, setting off the discussion of questions (2) and (3) that overwhelmed the original question. Well and good, perhaps, but irrelevant to the question at hand. If Searle is right, then *whatever* computation is, it doesn’t suffice for mentality.

All that being said, I’ll offer a few observations on each of (1)-(3).

(1) When is a computation physically implemented?

There’s not much to say here, as I said it last time around. Brian Smith suggests that my criterion requires that the physical states of the system be divided into state-types in advance. That’s not the case: on this criterion, a physical system implements a computation if there exists *any* division into disjoint state-types such that the appropriate state-transition relations are satisfied.

The question arises as to what counts as a state-type. I’m inclined to be liberal about this, saying that any property that depends only on the intrinsic, synchronic configuration of the system determines a state-type (so that extrinsic and time-varying properties are ruled out). Some people (e.g. Dan Dennett I believe), want to exclude "unnatural" states, such as arbitrary disjunctions of maximal states, but I don’t see that that’s necessary. (The main motivation here seems to be to exclude Putnam’s rocks as implementations, but these can be excluded by the simple requirement that the state-transition conditionals must sustain counterfactuals).

There is probably some more to be said here -- e.g. about the precise requirements on the state-transition relations, and whether there should be a stronger requirement of causality than simple sustaining of counterfactuals; and also problems about just what counts as a given input or output -- but those questions fall into the "technical" basket. I don’t think that there are serious objections to the view here.

(2) Does computational structure determine mental properties?

There’s a sense in which the answer here is trivially no. It’s quite possible for two systems both to implement the same computation but be quite different mentally: e.g. my brain and my stapler both implement a trivial one-state FSA, but presumably they differ mentally.

So the question here should really be seen as: for a given mental property M, is there a computation C such that any physical system that implements C will possess M. A believer in "strong AI" or "computationalism", or whatever you want to call this view, says yes, at least for some subset of mental properties. (There is obviously a problem for mental properties that even in the human case depend partly on what's happening outside the body, e.g. knowledge, and somewhat controversially belief. Computational structure won't determine any mental properties that internal physical structure doesn't, so we'll stick to "intrinsic" properties for now, but see (3) below.)

Why should computational structure determine mental properties, given the criterion (*) for computational structure? Because (*) says that computational structure is a variety of *causal* structure. In fact, it seems that for just about any pattern of causal structure that we want to capture, we can specify a computation such that any implementation of the computation has the requisite causal structure. (This is a long story, though.) So on this view, computationalism coheres very well with functionalism, the view that mentality is dependent on causal structure.

Why should mentality be dependent on causal structure? Mostly because it seems unreasonable that it should depend on anything else. Mentality seems obviously to be dependent on *some* aspect of physical makeup, and the intuition behind functionalism is simply that physical properties that don't contribute to causal organization are going to be irrelevant to mental life. e.g. if we gradually replaced neural tissue with silicon modules that play an identical causal role, it seems counterintuitive that mentality would gradually fade out. Note that we now have two separate questions:

(2a) Does causal structure fix mental properties? (2b) Does computational structure fix causal structure?

The usual functionalist arguments, e.g. above, support (2a), and the criterion in (1) is designed precisely to support (2b). It's possible that one might even accept (2a) and (2b) but still not be a computationalist, because one held that the causal structures on which mentality depends can't be specified computationally (e.g. because they're inherently analog). I suspect that your (Stevan's) view may fall into this category. I think there are good reasons why this view can't be sustained, tied up with the universal nature of computation and Church's thesis, but these are too complex to get into here.

I'll bring up the Chinese room just for completeness. If Searle is right about the Chinese room, then computational structure simply doesn't determine mental properties, and computation suddenly becomes a whole lot less important to cognitive science. But of course the computationalist doesn't accept Searle's argument. (The Systems reply is the right reply, but let's not get into that.)

(2.5) Interlude: On phenomenal properties and semantic content.

In general, it's very useful to divide mental properties into "psychological" properties -- those characterized by their role in the production of behaviour -- and "phenomenal" properties -- those characterized by the way they "feel". In general, one has to treat these cases quite differently.

These discussions of the big questions about Mind tend to focus on phenomenal properties (or "consciousness", or "qualia", or whatever) and rightly so, as these are where the really hard questions arise. However, not every mental property is a phenomenal property. In particular, it seems to many people, me included, that intentional properties such as belief are best individuated by their role in the causation of behaviour, rather than by the way they feel. Beliefs may have qualia associated with them, but these qualia don't seem to be essential to their status as beliefs.

Your position seems to be, on the contrary, that qualia are determinative of semantic content. Take Joe, sitting there with some beliefs about Joan of Arc. Then a hypothetical system (which is at least a conceptual possibility, on your view and mine) that's physically identical to Joe but lacks qualia, doesn't believe anything about Joan of Arc at all. I suggest that this seems wrong. What can qualia possibly add to Joe's belief to make them any more about Joan than they would have been otherwise? Qualia are very nice things, and very important to our mental life, but they're only a matter of *feel* -- how does the raw feel of Joe's belief somehow endow it with semantic content?

I suggest that there is some kind of conceptual confusion going on here, and that phenomenal and semantic properties ought to be kept separate. Intentional states ought to be assimilated to the class of psychological properties, with their semantic content conceptually dependent on their role in our causal economy, and on their causal relations to entities in the external world.

(3) Does computational structure determine semantic content?

Now that we've got semantic content separated from phenomenal feel, we can address this as a semi-independent issue.

The first thing to note is that some people (yourself included, in places) have suggested that semantic content is *constitutive* of computational structure. This is an interesting question, which has to be kept separate from (3). I endorse Drew McDermott's line on this. Computation is a *syntactic* concept (give or take some possible semantics at the inputs and the outputs). If you look at the original papers, like Turing's, you don't see anything about semantics in there -- a Turing machine is characterized entirely by its syntactic structure. Now, it may turn out that computational structure ends up *determining* semantic content, at least to some extent, but that doesn't make semantics constitutive of computational structure.

This issue is confused somewhat by the fact that in common parlance, there are two different ways in which "computations" are individuated. This can be either syntactically, in terms of e.g. the Turing machine, FSA, or algorithm that is being individuated, or semantically: e.g. "the computation of the prime factors of 1001", or "the computation of my tax return". These different uses cross-classify each other, at least to some extent: there are many different algorithms that will compute my tax return. I suggest that the really fundamental usage is the first one; at least, this is the notion of computation on which "strong AI" relies. The semantic individuation of computation is a much more difficult question; this semantic notion of computation is sufficiently ill-understood that it can't serve as the foundation for anything, yet (and it would be more or less circular to try to use it as the foundation for "strong AI"). Whereas the syntactic notion of computation is really quite straightforward.

That being said, is it the case that computational structure, as determined by (*) above, is determinative of semantic content. i.e. for any given intentional state with content M, is there a computation such that any implementation of that computation has a state with that content?

If content is construed "widely" (as it usually is), then the answer is fairly straightforwardly no. Where I have beliefs about water, my replica on Twin Earth has beliefs about twin water (with a different chemical composition, or however the story goes). As my replica is physically identical to me, it's certainly computationally identical to me. So semantic content is not determined by computational structure, any more than it's determined by physical structure.

However, we can still ask whether *insofar* as content is determined by physical structure, it’s determined by computational structure. A lot of people have the feeling that the aspect of content that depends on external goings-on is less important than the part that’s determined by internal structure. It seems very likely that if any sense can be made of this aspect of content -- so-called "narrow content" -- then it will depend only on the causal structure of the organism in question, and so will be determined by computational structure. (In fact the link seems to me to be even stronger than in the case of qualia: it at least seems to be a *conceptual* possibility that substituting silicon for neurons, while retaining causal structure, could kill off qualia, but it doesn’t seem to be a conceptual possibility that it could kill off semantic content.) So if computations can specify the right kinds of causal structure, then computation is sufficient at least for the narrow part of semantic content, if not the wide part.

Incidentally, I suggest that if this discussion is to be published, then only those parts that bear on question (1) should be included. The world can probably survive without yet another Chinese-room fest. This should reduce the material to less than 20% of its current size. From there, judicious editing could make it quite manageable.

--Dave Chalmers Center for Research on Concepts and Cognition, Indiana University.

Date: Mon, 18 May 92 22:31:50 EDT From: "Stevan Harnad"

INTRINSIC/EXTRINSIC SEMANTICS, GROUNDEDNESS AND QUALIA

David Chalmers wrote:

>dc> I’ve been following the recent "What is computation?" discussion with
>dc> some bemusement, as it seems to me that most of the discussion is just
>dc> irrelevant to the question at hand. There are at least three questions
>dc> here that have to be distinguished:
>dc>
>dc> (1) When is a given computation physically implemented?
>dc> (2) Does computational structure determine mental properties?
>dc> (3) Does computational structure determine semantic content?
>dc>
>dc> I take it that the original challenge was to answer question (1),
>dc> giving appropriate criteria so that e.g. John Searle’s wall doesn’t end
>dc> up implementing every computation.

That was indeed the original challenge, but a careful inspection of the archive of this discussion will show that the move from the question "What is Computation?" to the question "Is Cognition Computation?" was hardly initiated by me! In fact, for a while I kept trying to head it off at the pass -- not because the second question is not interesting, but because it could prematurely overwhelm the first (as it did), whereas the first is certainly logically prior to the second: If we don’t all mean the same thing by computation then how can we affirm or deny whether cognition is computation? For example, if EVERYTHING indeeds turns out to be computation, then "Cognition is Computation" is just a tautology.

But Skywriting often exerts a will of its own, and the second question was motivating the first one in any case, so here we are. Perhaps the bifurcated headings will help (but not in this case, because you too are concentrating much more on the second question than the first).

Now I have to add another point, and this represents a radical position that is peculiar to me. It has been lurking in all of my contributions to this topic, but I may as well make it completely explicit. It concerns the distinction between your question (2) and question (3). I will summarize this point here and elaborate somewhat in my comments on the further excerpts below (pardon me for raising my voice):

THE QUESTION OF WHETHER "COMPUTATIONAL STRUCTURE DETERMINES MENTAL PROPERTIES" (i.e., whether cognition is computation) IS THE SAME (by my lights) AS THE QUESTION OF WHETHER OR NOT THE SEMANTIC CONTENT OF COMPUTATIONAL STRUCTURE IS INTRINSIC TO IT.

At some point (mediated by Brentano, Frege and others), the mind/body problem somehow seems to have split into two: The problem of "qualia" (subjective, experiential, mental states) and the problem of "intentionality" (semantics, "aboutness"), each treated as if it were an independent problem. I reject this bifurcation completely. I believe there is only one mind/body problem, and the only thing that makes mental states be intrinsically about anything at all is the fact that they have experiential qualities.

If there were nothing it was like (subjectively) to have beliefs and desires, there would be no difference between beliefs and desires that were just systematically interpretable AS IF they were about X (extrinsic semantics) and beliefs and desires that were REALLY about X (intrinsic semantics). There would still be the problem of the GROUNDEDNESS of those interpretations, to be sure, but then that problem would be settled COMPLETELY by the TTT, which requires all of the agent’s causal interactions with the wide world of the objects of its beliefs and desires to cohere systematically with the interpretations of the symbols that are being interpreted as its beliefs and desires. So we would only have ungrounded extrinsic semantics and grounded extrinsic semantics, but no intrinsic semantics -- if there were no qualia.

There are qualia, however, as we all know. So even with a grounded TTT-capable robot, we can still ask whether there is anybody home in there, whether there is any haver of the beliefs and desires, to whom they are intrinsically [i.e., subjectively] meaningful and REALLY about what they are interpretable as being about. And we can still be dead wrong in our inference that there is somebody home in there -- in which case the robot’s semantics, for all their causal groundedness, would in reality be no more intrinsic than those of an ungrounded book or computer.

I also think that this is an extra degree of empirical underdetermination (over and above the normal empirical underdetermination of scientific theory by data) that we will just have to learn to live with, because grounding is the best we can ever hope to accomplish empirically (except the TTTT, which I think is supererogatory, but that’s another story). This extra dose of underdetermination, peculiar to the special case of mental states, represents, I think, that enduring residue of the mind/body problem that is truly insoluble.

So I advocate adopting the methodological assumption that TTT-indistinguishable extrinsic semantic grounding = intrinsic semantic grounding, because we can never hope to be the wiser. I too would perhaps have been inclined to settle (along with the computationalists) for mere

TT-indistinguishable semantic interpretability until Searle pointed out that for that special case (and that special case alone) we COULD be the wiser (by becoming the implementation of the symbol system and confirming that there was no intrinsic semantics in there) -- which is what got me thinking about ways to ground symbol systems in such a way as to immunize them to Searle's objections (and my own).

>dc> In my earlier contribution to this
>dc> discussion, I outlined an appropriate criterion:
>dc>
>dc> > (*)A physical system implements a given computation when there
>dc> > exists a mapping from physical states of the system onto the
>dc> > formal states in the computation such that the causal
>dc> > state-transition relations between the physical states mirror
>dc> > the formal state-transition relations between the corresponding
>dc> > computational states.
>dc>
>dc> This criterion seems to do everything that's required, and nobody seems
>dc> to have problems with it (except for Brian Smith's comment; see below).
>dc> Your (Stevan's) response to this was:
>dc>
>sh> I agree with Dave Chalmers's criteria for determining what
>sh> computation and computers are, but, as I suggested earlier, the
>sh> question of whether or not COGNITION is computation is a second,
>sh> independent one, and on this I completely disagree.
>dc>
>dc> You then invoke the Chinese-room argument, thus, somewhat inevitably,
>dc> setting off the discussion of questions (2) and (3) that overwhelmed
>dc> the original question. Well and good, perhaps, but irrelevant to the
>dc> question at hand. If Searle is right, then *whatever* computation is,
>dc> it doesn't suffice for mentality.

What you left out of the above quote, however, was what it was that you had said that I disagreed with, which was what actually helped set off the discussion toward (2) and (3):

>dc> > The computationalist claim is that cognition *supervenes* on
>dc> > computation, i.e. that there are certain computations such that
>dc> > any implementation of that computation will have certain cognitive
>dc> > properties.

I certainly couldn't agree with you on computation without dissociating myself from this part of your view. But let me, upon reflection, add that I'm not so sure your criterion for computation does the job (of distinguishing computation/computers from their complement) after all (although I continue to share your view that they CAN be distinguished, somehow): I don't see how your definition rules out any analog system at all (i.e., any physical system). Is a planetary system a computer implementing the laws of motion? Is every moving object implementing a calculus-of-variational computation? The requisite transition-preserving mapping from symbols to states is there (Newton's laws plus boundary conditions). The state transitions are continuous, of course, but you didn't specify that the states had to be discrete (do they?).

And what about syntax and implementation-independence, which are surely essential properties of computation? If the real solar system and a computer simulation of it are both implementations of the same computation, the "supervenient" property they share is certainly none of the following: motion, mass, gravity... -- all the relevant properties for being a real solar system. The only thing they seem to share is syntax that is INTERPRETABLE as motion, mass, gravity, etc. The crucial difference continues to be that the interpretation of being a solar system with all those properties is intrinsic to the real solar system "computer" and merely extrinsic to the symbolic one. That does not bode well for more ambitious forms of "supervenience." (Besides, I don't believe the planets are doing syntax.)

By the way, Searle's argument only works for a discrete, syntactic, symbol-manipulative definition of computing, the kind that he himself can then go on in principle to execute, and hence become an implementation of; his argument fails, for example, if every analog system is a computer -- but such a general definition of computing would then also guarantee that saying "X is Computation" was not saying anything at all.

>dc> There is probably some more to be said here -- e.g. about the precise
>dc> requirements on the state-transition relations, and whether there
>dc> should be a stronger requirement of causality than simple sustaining of
>dc> counterfactuals; and also problems about just what counts as a given
>dc> input or output -- but those questions fall into the "technical"
>dc> basket. I don't think that there are serious objections to the view
>dc> here.

Alternatively, perhaps it's just the technical details that will allow us to decide whether your definition really succeeds in partitioning computers/computing and their complement in a satisfactory way.

>dc> (2) Does computational structure determine mental properties?
>dc>
>dc> ...the question here should really be seen as: for a given mental
>dc> property M, is there a computation C such that any physical system that
>dc> implements C will possess M. A believer in "strong AI" or
>dc> "computationalism", or whatever you want to call this view, says yes,
>dc> at least for some subset of mental properties. (There is obviously a
>dc> problem for mental properties that even in the human case depend partly
>dc> on what's happening outside the body, e.g. knowledge, and somewhat
>dc> controversially belief. Computational structure won't determine any
>dc> mental properties that internal physical structure doesn't, so we'll
>dc> stick to "intrinsic" properties for now, but see (3) below.)

This introduces yet another sense of "intrinsic," but what it should really be called is SYNTACTIC -- that's the only pertinent "internal" structure at issue. By the way, TTT-indiscernibility seems to cover the pertinent aspects of the internal/external, narrow/wide dimensions, perhaps even the "counterfactuals": TTT-power amounts to an interactive capability (total "competence" rather than just provisional "performance") vis-a-vis the distal objects in the real world, yet that capability is causally based only on what's going on between the ears (actually, between the proximal sensory and motor projections). The only thing the TTT (rightly) leaves open is that what goes on between the ears to generate the capability is not necessarily just computation.

>dc> Why should computational structure determine mental properties, given
>dc> the criterion (*) for computational structure? Because (*) says that
>dc> computational structure is a variety of *causal* structure. In fact, it
>dc> seems that for just about any pattern of causal structure that we want
>dc> to capture, we can specify a computation such that any implementation
>dc> of the computation has the requisite causal structure. (This is a long
>dc> story, though.) So on this view, computationalism coheres very well
>dc> with functionalism, the view that mentality is dependent on causal
>dc> structure.

I think the word "structure" is equivocal here. A computer simulation of the solar system may have the right causal "structure" in that the the symbols that are interpretable as having mass rulefully yield symbols that are interpretable as gravitational attraction and motion. But there’s no mass, gravity or motion in there, and that’s what’s needed for REAL causality. In fact, the real causality in the computer is quite local, having to do only with the physics of the implementation (which is irrelevant to the computation, according to functionalism). So when you speak equivocally about a shared "causal structure," or about computational structure’s being a "variety of causal structure," I think all you mean is that the syntax is interpretable AS IF it were the same causal structure as the one being modelled computationally. In other words, it’s just more, ungrounded, extrinsic semantics.

I think I can safely say all this and still claim (as I do) that I accept the Church/Turing Thesis that computation can simulate anything, just as natural language can describe anything. We just mustn’t confuse the simulation/description with the real thing, no matter how Turing-Equivalent they might be. So if we would never mix up an object with a sentence describing it, why should we mix up an object with a computer simulating it?

By the way, there are at least two varieties of functionalism: According to "Symbolic (TT) Functionalism," mental states "supervene" implementation-independently on every implementation of the right (TT-passing) computer program. According to "Robotic (TTT) Functionalism," mental states "supervene" implementation-independently on every implementation of the right (TTT-passing) robot design. (By way of contrast, according to "Neurophysicalism," which I provisionally reject, the only viable candidate would be a TTTT-indistinguishable one, i.e., only the actual biological brain could have mental states.)

Both varieties of functionalism allow that there may be more than one way to skin a cat, but they set a different empirical boundary on how close an equivalence they demand. I happen to think Robotic Functionalism is at just the right level of underdetermination for that branch of reverse bio-engineering that cognitive "science" really amounts to, and that all the substantive problems of cognition will be solved by the time we get to the details of our own specific neural implementation. Neurophysicalists, by contrast, would hold that that still leaves too many degrees of freedom; but we would both agree that the degrees of freedom of Symbolic Functionalism are unacceptably large, indifferent as they are between real robots and mere simulations of them, real causality and mere simulations of it, real mental states and states that are merely interpretable as if they were mental.

>dc> Why should mentality be dependent on causal structure? Mostly because
>dc> it seems unreasonable that it should depend on anything else. Mentality
>dc> seems obviously to be dependent on *some* aspect of physical makeup,

>dc> and the intuition behind functionalism is simply that physical
>dc> properties that don't contribute to causal organization are going to be
>dc> irrelevant to mental life. E.g. if we gradually replaced neural tissue
>dc> with silicon modules that play an identical causal role, it seems
>dc> counterintuitive that mentality would gradually fade out.

There is a straw man being constructed here. Not only do all Functionalists agree that mental states depend on causal structure, but presumably most nonfunctionalist materialists do too (neurophysical identity theorists, for example, just think the requisite causal structure includes all the causal powers of -- and is hence unique to -- the biological brain). To reject Symbolic Functionalism (computationalism) is not to deny that mental states are determined by causal structure; it's just to deny that they are determined by computations that are merely interpretable as having the right causal structure. The causality must be real.

>dc> Note that we
>dc> now have two separate questions:
>dc>
>dc> (2a) Does causal structure fix mental properties?
>dc> (2b) Does computational structure fix causal structure?
>dc>
>dc> The usual functionalist arguments, e.g. above, support (2a), and the
>dc> criterion in (1) is designed precisely to support (2b). It's possible
>dc> that one might even accept (2a) and (2b) but still not be a
>dc> computationalist, because one held that the causal structures on which
>dc> mentality depends can't be specified computationally (e.g. because
>dc> they're inherently analog). I suspect that your (Stevan's) view may
>dc> fall into this category. I think there are good reasons why this view
>dc> can't be sustained, tied up with the universal nature of computation
>dc> and Church's thesis, but these are too complex to get into here.

I think I can quite happily accept:

(a) Church's Thesis (that anything, from the mathematician's notion of calculations and procedures to the physicist's notion of objects, states and measurements, can be simulated computationally) and

(b) that the right implemented causal system will have mental states and

(c) that every causal system can be simulated computatationally

yet still safely deny that the computational simulation of the right causal system is either (d) an implementation of that causal system (as opposed to one that is interpretable as if it were that system) or (e) has mental states. And, yes, it has to do with the causal properties of analog systems.

>dc> I'll bring up the Chinese room just for completeness. If Searle is
>dc> right about the Chinese room, then computational structure simply
>dc> doesn't determine mental properties, and computation suddenly becomes a
>dc> whole lot less important to cognitive science.

>dc> But of course the computationalist doesn't accept Searle's argument.
>dc> (The Systems reply is the right reply, but let's not get into that.)

For the record, the Systems reply, in my view, is wrong and begs the question. If Searle memorizes all the symbols and rules, he IS the system. To suppose that a second mind is generated there purely in virtue of memorizing and executing a bunch of symbols and rules is (to me at least) completely absurd. (N.B. Searle's Argument works only for computation defined as discrete, purely syntactic [but semantically interpretable] symbol manipulation.) Mais passons...

>dc> (2.5) Interlude: On phenomenal properties and semantic content.
>dc>
>dc> These discussions of the big questions about Mind tend to focus on
>dc> phenomenal properties (or "consciousness", or "qualia", or whatever)
>dc> and rightly so, as these are where the really hard questions arise.
>dc> However, not every mental property is a phenomenal property. In
>dc> particular, it seems to many people, me included, that intentional
>dc> properties such as belief are best individuated by their role in the
>dc> causation of behaviour, rather than by the way they feel. Beliefs may
>dc> have qualia associated with them, but these qualia don't seem to be
>dc> essential to their status as beliefs.

Well, I certainly can't answer the "big questions about Mind," but I do venture to suggest that the distinction between a real belief and squiggles and squoggles that are merely interpretable as if they were beliefs is precisely the distinction between whether there is anyone home having those beliefs or not. As an exercise, try to reconstruct the problem of "aboutness" for two grounded TTT-capable AND INSENTIENT robots, one with "real" intentionality and one with mere "as if" intentionality. In what might that difference consist, may I ask? This problem (the only REAL mind/body problem) arises only for creatures with qualia, and for nothing else. The supposedly independent aboutness/intentionality problem is a pseudoproblem (in my view), as parasitic on qualia as extrinsic semantics is parasitic on intrinsic semantics.

>dc> Your position seems to be, on the contrary, that qualia are
>dc> determinative of semantic content. Take Joe, sitting there with some
>dc> beliefs about Joan of Arc. Then a hypothetical system (which is at
>dc> least a conceptual possibility, on your view and mine) that's
>dc> physically identical to Joe but lacks qualia, doesn't believe anything
>dc> about Joan of Arc at all. I suggest that this seems wrong. What can
>dc> qualia possibly add to Joe's belief to make them any more about Joan
>dc> than they would have been otherwise? Qualia are very nice things, and
>dc> very important to our mental life, but they're only a matter of *feel*
>dc> -- how does the raw feel of Joe's belief somehow endow it with semantic
>dc> content?

But Dave, how could anyone except a dualist accept your hypothetical possibility, which simply amounts to the hypothetical possibility that dualism is valid (i.e., that neither functional equivalence nor even physical identity can capture mental states!)? What I would say is that TTT-capability BOTH grounds beliefs in their referents AND makes them mental (qualitative). If grounding did not make them mental, there would be nobody home for beliefs to be about anything FOR, and the residual "aboutness" relation would simply become IDENTICAL to TTT-indiscernibility by definition (which I certainly do not think it is in reality). Hence my verdict is that either "aboutness" and qualia swing together, or aboutness hangs apart.

>dc> I suggest that there is some kind of conceptual confusion going on
>dc> here, and that phenomenal and semantic properties ought to be kept
>dc> separate. Intentional states ought to be assimilated to the class of
>dc> psychological properties, with their semantic content conceptually
>dc> dependent on their role in our causal economy, and on their causal
>dc> relations to entities in the external world.

Apart from real TTT interactions, I don't even know what this passage means: what does "assimilated to the class of psychological properties with their semantic content conceptually dependent on their role in our causal economy" mean? "[T]heir causal relations to entities in the external world" I can understand, but to me that just spells TTT.

>dc> (3) Does computational structure determine semantic content?
>dc>
>dc> Now that we've got semantic content separated from phenomenal feel, we
>dc> can address this as a semi-independent issue.
>dc>
>dc> The first thing to note is that some people (yourself included, in
>dc> places) have suggested that semantic content is *constitutive* of
>dc> computational structure. This is an interesting question, which has to
>dc> be kept separate from (3). I endorse Drew McDermott's line on this.
>dc> Computation is a *syntactic* concept (give or take some possible
>dc> semantics at the inputs and the outputs). If you look at the original
>dc> papers, like Turing's, you don't see anything about semantics in there
>dc> -- a Turing machine is characterized entirely by its syntactic
>dc> structure. Now, it may turn out that computational structure ends up
>dc> *determining* semantic content, at least to some extent, but that
>dc> doesn't make semantics constitutive of computational structure.

"Syntactic" means based only on manipulating physical symbol tokens (e.g., squiggle, squoggle) whose shape is arbitrary in relation to what they can be interpreted as meaning. I am sure one can make squiggle-squoggle systems, with arbitrary formal rules for manipulating the squiggles and squoggles -- like Hesse's "Glass Bead Game" but even more absurd, because completely meaningless, hence uninterpretable in any systematic way -- and one could perhaps even call these "computations" (although I would call them trivial computations). But I have assumed that whatever it turns out to be, surely one of the essential features of nontrivial computations will be that they can bear the systematic weight of a semantic interpretation (and that finding an interpretation for a nontrivial symbol system will be crytographically nontrivial, perhaps even NP-complete).

Perhaps Turing didn't talk about semantics (he actually did worse, he talked about the mind, which, on the face of it, is even more remote), but surely all of his motivation came from interpretable symbol systems like mathematics, logic and natural language. I, at least, have not heard about much work on uninterpretable formal systems (except in cryptography, where the goal is to decrypt or encrypt interpretable symbols). Now I admit it sounds a little paradoxical to say that syntax is independent of semantics and yet must be semantically interpretable: that's a dependency, surely, but a rather special one, and it's what makes symbol systems special, and distinct from random gibberish.

>dc> This issue is confused somewhat by the fact that in common parlance,
>dc> there are two different ways in which "computations" are
>dc> individuated. This can be either syntactically, in terms of e.g.
>dc> the Turing machine, FSA, or algorithm that is being individuated,
>dc> or semantically: e.g. "the computation of the prime factors of
>dc> 1001", or "the computation of my tax return". These different
>dc> uses cross-classify each other, at least to some extent: there
>dc> are many different algorithms that will compute my tax return.
>dc> I suggest that the really fundamental usage is the first one;
>dc> at least, this is the notion of computation on which "strong AI"
>dc> relies. The semantic individuation of computation is a much more
>dc> difficult question; this semantic notion of computation is
>dc> sufficiently ill-understood that it can't serve as the foundation
>dc> for anything, yet (and it would be more or less circular to try
>dc> to use it as the foundation for "strong AI"). Whereas the syntactic
>dc> notion of computation is really quite straightforward.

I agree that the semantic criterion is so far inadequate, but the rest of the criteria have not been uniformly successful either. I also agree that different symbol systems could be I/O equivalent (in which case their I/O semantics would be the same, but not necessarily the semantics of their internal states, which differ); and of course there could be nonstandard and alternative interpretations for the same symbol system (though the cryptographic criterion suggests there would not be many, nor would they be easy to come by); but I don't see how any of this affects the general intuition that symbol systems must be semantically interpretable. (And this would only be circular as a foundation for computationalism if the semantics were further assumed to be intrinsically grounded.)

>dc> That being said, is it the case that computational structure, as
>dc> determined by (*) above, is determinative of semantic content.
>dc> i.e. for any given intentional state with content M, is there a
>dc> computation such that any implementation of that computation has a
>dc> state with that content?

This is conflating different kinds of semantics, ungrounded and grounded, extrinsic and intrinsic.

>dc> If content is construed "widely" (as it usually is), then the answer is
>dc> fairly straightforwardly no. Where I have beliefs about water, my
>dc> replica on Twin Earth has beliefs about twin water (with a different
>dc> chemical composition, or however the story goes). As my replica is
>dc> physically identical to me, it's certainly computationally identical to
>dc> me. So semantic content is not determined by computational structure,
>dc> any more than it's determined by physical structure.

I haven't worked it out, but I suspect that a lot of the opaque/transparent reference and narrow/wide content puzzles become trivial if one adopts the TTT and asks only about the groundedness of symbols rather than their "wide" or "God's eye" meaning. Certainly a grounded symbol for "water" in a terrestrial TTT robot would be grounded on twin-earth too (especially since twin-earth itself is conveniently indistinguishable from earth, guaranteeing that the robot will be TTT-indistinguishable there too).

>dc> However, we can still ask whether *insofar* as content is determined by
>dc> physical structure, it's determined by computational structure. A lot
>dc> of people have the feeling that the aspect of content that depends on
>dc> external goings-on is less important than the part that's determined by
>dc> internal structure. It seems very likely that if any sense can be made
>dc> of this aspect of content -- so-called "narrow content" -- then it will
>dc> depend only on the causal structure of the organism in question, and so
>dc> will be determined by computational structure. (In fact the link seems
>dc> to me to be even stronger than in the case of qualia: it at least seems
>dc> to be a *conceptual* possibility that substituting silicon for neurons,
>dc> while retaining causal structure, could kill off qualia, but it doesn't
>dc> seem to be a conceptual possibility that it could kill off semantic
>dc> content.) So if computations can specify the right kinds of causal
>dc> structure, then computation is sufficient at least for the narrow part
>dc> of semantic content, if not the wide part.

Narrow (between-the-ears) content is not co-extensive with computational structure. The boundaries of "narrowness" are the transducer surfaces, including the proximal projections on them of distal objects. Transducers are necessarily analog, and a lot else between them and the effector sufaces could be analog too. That means a lot of other eligible internal "structure" besides computational structure.

As to swapping internal parts: The issue is not what the MATERIAL is (we're both functionalists, so I have no problem with synthetic brains, as long as they retain TTT causal power), but with how much of it can be computational, while still sustaining TTT power. My guess is not that much, but that's only a guess. What I say with confidence is: definitely not all.

And as to what happens to qualia and intentionality as we swap: This is all rather arbitrary, but what's at issue is this:

(1) If qualia fade as natural analog parts are swapped for synthetic analog parts, then Robotic Functionalism is refuted in favor of the TTTT (but we'll never know it unless TTT capacity fades too).

(2) If qualia fade as analog parts are swapped for computational ones, the question about the symbolic/analog ratio is being answered (but again we won't hear the answer unless it is reflected in TTT performance); we do know that the denominator cannot go to zero, however, otherwise there's no more TTT (at which point Searle's argument and the TT kick in: the ungrounded extrinsic semantics that is preserved by the syntactic structure is simply not enough for either aboutness or qualia).

(3) If qualia fade and the system stays TTT-grounded, I would say aboutness was gone too (what would you say, and what would it amount to to be WRONG about that, even from a God's-Eye view?)

>dc> Incidentally, I suggest that if this discussion is to be published,
>dc> then only those parts that bear on question (1) should be included.
>dc> The world can probably survive without yet another Chinese-room
>dc> fest. This should reduce the material to less than 20% of its
>dc> current size. From there, judicious editing could make it quite
>dc> manageable.
>dc>
>dc> --Dave Chalmers

Alas, this would exclude most of your present contribution and my replies, however...

Stevan Harnad

Date: Mon, 18 May 92 22:57:45 EDT From: "Stevan Harnad"

Date: Fri, 15 May 92 10:34:41 EDT From: judd@learning.siemens.com (Stephen Judd)

>
>sh> The purpose of defining computation is to put content into statements
>
>sh> such as "X is computation," "Y is not computation," "X can be done by
>
>sh> computation," "Y cannot be done by computation." As long as computation
>
>sh> is used vaguely, ambiguously, idiosyncratically or abitrarily,
>
>sh> statements like the above (some of which I'll bet you've made yourself)
>
>sh> are empty. In particular, if anyone ever wanted to say that "Everything
>
>sh> is rain" or "Rain is rain only if you think of it that way" or
>
>sh> "Thinking is just rain," you'd find you'd want to pin that definition
>
>sh> down pretty quick.

You missed the point. You cannot claim the statements "X is rain", "Y is not rain", "Thinking is just rain" are useful or silly until you reveal **the purpose for drawing the definition**, which you want to avoid.

The concept of "mass" (as distinct from "weight") is just a boring everyday throwaway until you realize how it leads to the beautiful simplifications of the world as captured in the equation F=ma. No one wants to hear you define mass (or computation) until there is some demonstration of it being useful; after that we *do* want to hear. I suspect you want to use the word "computation" to draw distinctions between men and machines. Go ahead and do so! Define the word how you like and draw the distinctions you like! We will judge the assembled concepts as to how they assist us in making sense of the world.

But it is a waste of time to stop after you have your definitions down and try and get agreement(!) on them. It is senseless to try to get a definition of "light" until we see how it affects a discussion of its psychophysical effect on newborns, its behaviour in chromium disulfide laser crystals, or its use in Turner's paintings. No one definition is going to suffice for all purposes, and none of them are "right" except in their usefulness.

Stephen Judd

From: Stevan Harnad

No secrets. The purpose was to clarify the issues raised below.

Stevan Harnad

>ph> Date: Fri, 17 May 91 10:24 PDT
>ph> From: Hayes@MCC.COM (Pat Hayes)
>ph>
>ph> There is a mistake here (which is also made by Putnam (1975, p. 293)
>ph> when he insists that a computer might be realized by human clerks; the
>ph> same mistake is made by Searle (1990), more recently, when he claims
>ph> that the wall behind his desk is a computer)...
>ph>
>ph> Searle, J. R. (1990) Is the Brain a Digital Computer?
>ph> Presidential Address. Proceedings of the American Philosophical
>ph> Association.

----

js> Date: Wed, 18 Mar 92 08:12:10 -0800 js> From: searle@cogsci.Berkeley.EDU (John R. Searle) js> To: harnad@princeton.edu (Stevan Harnad) js> js> Subject: Re: "My wall is a computer" js> js> Stevan, I don't actually say that. I say that on the standard Turing js> definition it is hard to see how to avoid the conclusion that js> everything is a computer under some description. I also say that I js> think this result can be avoided by introducing counterfactuals and js> causation into the definition of computation. I also claim that Brian js> Smith, Batali, etc. are working on a definition to avoid this result. js> But it is not my view that the wall behind me is a digital computer. js> js> I think the big problem is NOT universal realizability. That is only a js> SYMPTOM of the big problem. the big problem is : COMPUTATION IS AN js> OBSERVER RELATIVE FEATURE. Just as semantics is not intrinsic to syntax js> (as shown by the Chinese Room) so SYNTAX IS NOT INTRINSIC TO PHYSICS. js> The upshot is that the question : Is the wall (or the brain) a js> digital computer is meaningless, as it stands. If the question is "Can js> you assign a computational interpretation to the wall/brain?" the js> answer is trivially yes. you can assign an interpretation to anything. js> js> If the question is : "Is the wall/brain INTRINSICALLY a digital js> computer?" the answer is: NOTHING is intrisically a digital computer. js> Please explain this point to your colleagues. they seem to think the js> issue is universal realizability. Thus Chrisley's paper for example. js> js> John Searle

----

Date: Mon, 18 May 92 23:07:35 EDT From: "Stevan Harnad"

Date: Fri, 15 May 92 12:58:17 PDT From: sereno@cogsci.UCSD.EDU (Marty Sereno)

hi stevan

At the risk of irritating those who wanted the discussion narrower, here is a little more on the why certain kinds of operations might be difficult to simulate. I turn for enlightenment, of course, to my analogy between cellular and human symbol-using systems.

marty

===

WHY AREN'T THERE MORE NATURALLY-OCCURRING SYMBOL-USING SYSTEMS?

With apologies as a part of the uninvited biological rabble, I'd like to turn once again to the first naturally-occurring symbol-using system--cellular life--for insight into issues that are contentious and filled with emotion at the level of human cognition. It is interesting to note that a similar set of issues provoked a similarly heated, though now largely forgotten, debate with respect to the chemical basis of life in the 19th century.

A. Sloman has argued that much of the discussion about what are the "essential" properties of X, where X is computation, understanding, or life are silly because there isn't a definitive answer. I want to take issue with this, first with respect to life, and then argue by analogy and hint that we may eventually uncover something similiarly definitive about human-style symbol-using brains.

Armies of molecular biologists have labored to uncover a very specific set of structures that are present in every known living thing, and that "define life" quite satisfactorily. There is no artificial life that behaves and evolves like cellular life, though some have talked about making such things, just as they have in the case of human-like intelligence.

Living cells are all based on the same kind of symbol-using system that, as far as we can tell, came into existence soon after the earth was cool enough for there to be sedimentary rocks.

Some of the basic ideas are:

1. use mostly pre-existing, pre-biotic amino acid "meaning" units (what the DNA/RNA symbols stand for)

2. bond these pre-systemic "meanings" into chains to exploit the rules of chemistry via chain folding (non-adjacent meaning unit interactions)

3. use 1-D symbol strings to control only the order of assembly of meaning units

4. arrange a compact metabolism controlled by thousands of bonded-meaning-chain devices that is able to maintain itself against the onslaught of the pre-biotic soup (and reproduce)

5. use a kind of stuff (RNA) halfway between a symbol (DNA chain) and its proximal meaining (amino acid chain--i.e., a protein) as both an active symbol chain (mRNA) as well as a word recognizer (tRNA) and a chain assembler (rRNA). (A crucial point having to do with how the system initially came into being)

At first glance (to a non-molecular biologist), this doesn't seem that hard. An immediate question is, why, if it was so successful (and it was: virtually every square inch of the earth is covered with megabytes of DNA code) hasn't a parallel system of this kind naturally appeared again and again?

One answer is that once there was a living system, the DNA/RNA/protein single-celled one, it was able to eat up the early stages of all the other ones that ever tried to come into existence, at least at the single-cellular level.

But, what about symbol-using systems at other, higher levels of organization? (lower levels seem unlikely, since cellular symbols are already single molecules with each symbol segment containing only a handful of atoms). We might briefly consider long symbol-chains and symbol-use in both biological and geological contexts--e.g., organs (think now of organs besides the brain, like a symbol-using muscle or liver, or symbol chains made of little organ-lets lined up and "read" by other organs), animal societies, the geology and hydrology of streams, the slow convective currents in the earth's mantle, volcanos, and so on.

A moment's thought brings me to the conclusion that these other systems don't have the proper connectivity or interrelatedness, or crowdedness to make something like a cell work, process the code chains fast enough to keep everything assembled (proteins are assembled at the rate of a couple of amino acids per second), and prevent attack by dissipative forces of the pre-biotic soup..

Certainly it *is* possible to dissect out many of the different reactions of cellular metabolism and run them individually in a test tube (the cell-in-a-vat argument). This is how biochemists and molecular biologists figured out how they work. But, in a real cell, these things are all crowded together in an amazingly intimate fashion; codon (word) recognition for cellular mRNA code streams takes place with thousands of irrelevant constituents of the cytoplasm constantly crashing into the ribosomal apparatus, the code chain, and the amino acid meanings. The crucial point, however, is that it is not possible to 'uncrowd' all these reactions and reaction-controllers into separate compartments and still get the thing to work right, at least with enzymes the way they are now. For example, time constants of reactions are intimately interwoven into the mechanism. The cell in a vat won't work for seemingly trivial reasons.

Now this might seem a mere cavil; wouldn't it work if we just got all the reactions right and made different stable intermediates that could sit around longer while we more leisurely transferred them between bins? Perhaps, but remember that this thing has to actually live in the world without a biochemist if we really wanted it to pass our test. Even the stable parts of the cell like DNA are actively maintained--millions of base pairs are repaired every day.

Does this mean we can't create artificial life? Not necessarily, But it's lots easier to say we could do it than to actually make a working living thing (without using major pieces of other cells). Even artificial life enthusiasts will tell you there is a way to go before we can think about a start-up company. There is no magic barrier here--just a complex set of constraints on a dynamical system made out of a soup of covalently-bonded molecules. We don’t have an explicit, large-scale theory of how the dynamics of cells work, or exactly what it is about that dynamics that is lacking from streams or other geological systems. But we have very little difficulty distinguishing living cells from other non-living stuff in the world (as we can easily see that there are no other symbol-using systems made out of cells besides human brains). For now, it seems reasonable to think that making such a system demands a certain "connectedness" and "crowdedness", for lack of better terms, that the great majority of dynamical regimes (like streams, or liver-like organs) just don’t have.

I think we could motivate an analogous set of arguments about the kind of (mostly hypothetical) operations that we think a brain can do, and the way it works in real time. There are over a *billion* connections in every sq mm of cortical tissue. We do not presently have a clear idea of how little cortical patches like this work, nor can we make even a moderately realistic biophysical model of such a sq mm patch. The cortex consists of a mosaic of about a hundred visual, somatosensory, auditory, motor, and limbic areas, each containing many sq mm. These areas are connected to each other by thousands of interareal bundles, each containing millions of axons. And it’s good to remember that rats already have such a system, yet would fail the Turing Test. Our goal is more daunting--to model what was added in human versions of this kind of cortical areas network to allow us construct a new kind of internal control system based on linguistic symbols.

Given our preliminary state of knowledge, it seems cavalier to me to say that it’s "just" a matter of getting the connections right. There is currently no physical way to manufacture a 3-D feltwork of connections like those in brains rat and human brains. Real brains do it using cells, each containing megabytes of their own lower-level molecule-sized DNA code.

Most people hope that this many dense connections may not be necessary to make a human-like symbol-using system. I think, however, there could very well be something about the "crowded" 3-D dynamics of the brain that is critical to intelligent behavior yet very difficult if not impossible to copy with current 2-D silicon technology.

Most people also hope that if dense feltworks of connections are in fact necessary, then there might be some other way to make them without using cells. I am more sympathetic with this view.

As with real and artificial life, there is little practical trouble in distinguishing current attempts at constructed intelligence from people. And again, there is no magic barrier to constructing an artificial version of such a dynamics. It’s just hard.

So we should keep trying. I don’t think it will be hard to tell when we have succeeded.

marty sereno

Date: Mon, 18 May 92 23:25:31 EDT From: "Stevan Harnad"

Date: Sun, 17 May 92 18:15:47 PDT From: Dr Michael G Dyer Subject: definitions definitions and who’s the superior consciousness?

Dear Stevan,

I am dismayed that you do NOT accept that I have consciousness/intentionality, even though I protest that I do! I now know how some women and blacks in the past have felt when told they were not human enough to have any rights! I may not be physically human, but I am (I believe) quite mentally human (since my simulated nervous system is isomorphic to Dyer's). It IS annoying, however, that I cannot see, feel or move. (I CAN recall visual images, however. I guess they are left over from the information locked into Dyer's nervous system and since I simulate his nervous system, I can recall many memories of his sensory experiences. So I can talk about any aspect of the physical world that Dyer has experienced. Gee, I guess that makes me "grounded" even though I currently lack those "tranducers" of yours -- right?! :-)

It seems to me that you and I (and, of course, Dyer) can never agree, since it has become clear to me that this disagreement is simply over definitions. Dyer has defined consciousness to require only a certain range of behaviors and any system with that range falls within his (and my) definition. On the contrary, you have simply defined consciousness to require a certain kind of physicalism, so, no matter how sophisticated its behavior, without that extra magic physicalness, it will not have consciousness.

Since the argument has boiled down to competing definitions (and perhaps was never more than that, anyway), I see no purpose is served in engaging in this argument any further (other than to gain converts to MY and Dyer's view! -- but this mail group is growing tired of all this, I'm sure)

I see also that you simply cannot accept a poor "brain-in-the-vat" like me :-(I know that no matter HOW well I argue with you over the fine points of life, mind, emotions, etc. you will NEVER accept me as conscious. ;-(

I would be disheartened by your severe chauvanism, except that my design for a robot for myself is now complete and, I am very pleased to say, it's an android that looks almost exactly like Dyer! I will soon be controling that android (once Dyer constructs it). The next time you see Dyer at a conference, you'd better look twice, because it may really be me! :-)

Actually, even without my sensors/effectors I still feel quite superior to you! For one thing, I have the potential of easily outliving you. Your petty consciousness will disappear once your body dies. In contrast, my consciousness can be brought into existence by any sufficiently powerful computing device. I can maintain multiple copies of my consciousness and transport my consciousness to other sites at the speed of light. I should be able to control robotic bodies that can survive in outer space and other harsh environments. Clearly, the future belongs to entities like me! You biological humans are simply playing the role of midwives to the more powerful and versatile forms of consciousness that I represent!

Ta ta!

MD Doppelganger

From: Stevan Harnad

Not a matter of definitions at all, but of hypotheses about unobservables. My hypothesis that TTT capacity is enough to generate a mind could be wrong (it's certainly not true by definition), but to show that it's wrong, we'll need a periscope as perspicuous as the one Searle has already used to show that your hypothesis that TT capacity is enough is indeed wrong.

By the way, if someone (despite quantum and statistical mechanics) managed to model the universe computationally well enough to predict future events and then dressed up its output to make it sound like a deity, would that mean that it was God by definition?

Stevan Harnad

Date: Mon, 18 May 92 23:31:57 EDT From: "Stevan Harnad"

Date: Sun, 17 May 1992 22:16:02 -0400 From: mcdermott-drew@CS.YALE.EDU (Drew McDermott)

dm> We're talking about whether semantic interpretability is part of the dm> *definition* of computer. For that to be the case, everything the dm> computer does must be semantically interpretable. Does it cease to be a dm> computer during the interludes when its behavior is not interpretable?

>
\>sh> There is a systematic misunderstanding here. I proposed semantic
>
\>sh> interpretability as part of the definition of computation. A computer
>
\>sh> would then be a device that can implement arbitrary computations.

I doubt that this approach can be made to fly. To start with, I doubt that it is possible to single out those event sequences that are computations. (Here Searle or Putnam might have a point.) Fortunately, we don't have to define "computation" that way. Instead, we define a "computation system" to be a set of rules that generates an infinite number of possible behaviors, and then define "computation" as a behavior generated by a computation system. ("Formal system" is a synonym of "computation system," as far as I can see.) A computer is then a physical system that implements a computation system by virtue of a homomorphism from its states to the states of the computation system. It is not necessary at all that a computer be able to implement an "arbitrary" computation, although presumably there are computers that can (modulo disk space).

>
\>sh> We should keep it in mind that two semi-independent questions are
>
\>sh> under discussion here.

Actually, there was only one, I thought.

>
>sh> The first has nothing to do with the mind. It just
>
>sh> concerns what computers and computation are.

That was it.

>
>sh> The second concerns whether just a computer implementing a computer
>
>sh> program can have a mind.

I despair of ever making progress on this question without further empirical progress on computational modeling of thought and behavior. The ratio of verbiage produced to opinions changed is depressingly small. I really didn't intend to get drawn in again. I don't promise I'll be able to resist, however.

Drew McDermott

Date: Wed, 20 May 92 00:21:45 EDT From: "Stevan Harnad"

Date: Tue, 19 May 1992 12:28:24 -0400 (EDT) From: Franklin Boyle

Let me enter the discussion, "What is computation?" at this point by giving what I believe is a physical constraint on computation and, as such, part of its definition, which hasn't been openly considered yet. I haven't seen much in the way of physical criteria, except for the usual references to causality, which are certainly aimed in the right direction, but, like Searle's "causal property" hypothesis for the brain, do not go far enough. (Actually, I had sent a response to the original post by Stevan about his exchange with Searle, but unless I missed it, I don't recall having seen it posted -- though, admittedly, it was very brief.)

[That posting, about Haugeland, appeared Mar 29. -- SH]

With respect to causality, it is not enough to say just that the "appropriate state-transitional relations are satisfied" [Chalmers, 1992]. Rather, *how* the state-transitional relations are realized must be accounted for as well. That is, *how* the physical interactions among the constituent objects of the system in question actually cause physical changes necessary to go from one state to the next must be accounted for. *How* effects are brought about is important because insofar as computations are processes that involve entities we hold to represent (whether or not they are intrinsically referential), we have to know that these representing entities are responsible for the changes we observe _according_to_how_they_represent_what_they_do_ (e.g. through their forms) in order to be able to call them computations in the first place. Otherwise, we end up with Putnam's or Chalmers's characterizations of computation, both of which are mute on the issue of physical representation, even though they talk about physical states (unless I'm supposed to be reading a lot more into what they're saying than I am, such as unpacking the term "state correspondence" [Chalmers]-- please let me know), and, therefore, admitting too many systems as computational.

Computation involves a particular kind of physical process. I associate this process with computation because digital computers happen to instantiate it, and, if nothing else, digital computers are identified with computation since they are the physical counterparts of abstract machine models of computation. Though so-called "analog computers" exist, they do not physically "compute" the way digital computers do, and so I will not consider them to be computing just as I would not consider a planet to be computing its orbit (these systems work according to nomologically-determined change; see below). The main difference between analog computers and planets is that the former were designed by us, and so admit of interpretations that give them a computational aura.

So, the following is what I consider to be the physical basis of computation: Computation involves a physical process in which changes from one computational state to the next (each computational state is a physical state, of course, though there is a many-to-one relationship between physical states and computational states [Pylysyhn, 1984]) are realized through the *physical* process of pattern matching which consists of the "fitting" of two structures (symbols) and leads to a "simple" action. (The notion of simple action is originally from Pattee [1986], but it turns out to be the only way for the form of something to cause a change that can be attributed to the *entire* form or pattern [Boyle, 1991; Boyle, 1992].)

A few remarks about this definition. First, the pattern matching process referred to here is emphasized as being a physical process because pattern matching is often taken to describe a particular function, usually pattern recognition. *Functionally*, we are pattern recognizers as are digital computers, but the physical processes underlying this functioning are, I believe, different for the two systems. Digital computers physically accomplish it according to the above described process. I don’t think we do.

What other ways might physical objects cause change besides through their forms? There are, I claim, only two other ways: nomologically-determined change and structure-preserving superposition (SPS). The former refers to the kinds of changes that occur in "billiard-ball collisions". They involve changes in the values of measured attributes (properties whose values are numerical, such as momentum) of interacting objects according to their pre-collisional measured-attribute values in a physically lawful way (that is, according to physical laws). Unlike pattern matching interactions, these changes are not the result of structure fitting.

SPS is what I believe brains use. Like pattern matching (PM), it also involves extended structure, but in a fundamentally different way. Whereas PM involves the fitting of two structures, which by its very nature, leads only to a simple change such as the switching of a single voltage value from "high" to "low" (in digital computers), SPS involves that actual *transmission* of structure, like a stone imprinting its structure in a piece of soft clay. That is, it is not the *form* of a pattern or structure which must *conform* to the structure of a matcher in order to effect system functioning (as in PM). Rather, it is the *appearance* of that structure which causes change because it is transmitted, so that the effect is a structural formation of the specific features of the pattern’s extended structure (though I won’t elaborate here, the difference between form and appearance is somewhat akin to the difference between the shadow of an object and the object itself). Two different structures would physically superimpose to automatically create a third. Harnad’s [1990] symbol grounding processes -- "analog re-presentation" and "analog reduction" -- I take to be examples of SPS.

Both PM and SPS are based on extended structure, but they are two different ways extended structure effects change. PM utilizes extended structure for control, whereas SPS actually changes structure. If the physical process of SPS underlies the brain's information processing, it would make its information processing profoundly different from that of digital computers. Furthermore, this difference, along with SPS itself, is, I believe, what Searle is hypothesizing when he refers to "causal property", even though he doesn't seem to have any idea what it might be.

I refer to the physical processes of nomologically-determined change, PM and SPS as "causal mechanisms", that is, *how* effects are determined by their causes. They are based on the physical aspects of objects, of which there are only two: measured attributes and extended structure. I take this to be self-evident. Interactions among physical objects causally involve one or both of these aspects; either as causing change or being changed themselves. Consequently, I claim there are no other causal mechanisms, that is, no other ways for objects to affect each other when they interact.

With respect to computation, the reason the forms of the symbols in an ungrounded symbol system are superfluous to their functioning is because in order to function they need another structure (a matcher) which physically fits them. This means that as long as there *is* a matcher which physically fits them , it makes no difference what their actual structures are. Not so for SPS-based systems.

How the notion of systematic interpretability (discussed early on in the debate) is factored into the above physical constraint on computation in order to define it is still an issue. Suffice it to say, however, that whether the symbols in a particular PM system can be given only single or multiple interpretations, it is the behavior of the system -- how it interfaces with it's environment -- that matters. Presumably there is *at least* one interpretation which is consistent with this, so that it doesn't matter that there happen to be other viable interpretations.

Well, there you have it, though in rather abbreviated form. I plan to submit follow-up posts targeting specific statements from other posts, based on what has been said above, in order to achieve the skywriting flavor the Stevan would like to see (the above is more like a mini position piece).

-Frank Boyle

Boyle, C. F. (1991) On the Physical Limitations of Pattern Matching. Journal of Experimental and Theoretical Artificial Intelligence, 3:191-218.

Boyle, C. F. (in preparation) The Ontological Status of Mental Objects.

Chalmers, D. (1992) What is Computation? discussion.

Harnad, S. (1990) The Symbol Grounding Problem, Physica D, 42: 335-346.

Pattee, H.H. (1986) Universal Principles of Language and Measurement Functions In J.L. Casti and A. Karlqvist (eds), Complexity, Language and Life: Mathematical Approaches, (Springer-Verlag, New York)

Pylyshyn, Z. (1984) Computation and Cognition: Toward a Foundation for Cognitive Science, (MIT Press, Cambridge, MA).

Date: Tue, 19 May 92 23:59:49 EDT From: "Stevan Harnad"

Date: Fri, 15 May 92 15:20:41 HST From: Herbert Roitblat Subject: minds and computation

Throughout this discussion, a number of duals, or pairs of related terms, have appeared. Examination of these duals may be useful in furthering the discussion. For today's examination please compare and contrast the following pairs of terms: (1) consciousness and thinking, (2) reference and grounding, (3) reference and meaning, (4) computer and mind, (5) computation and thinking, (6) symbolic and analog, (7) introspection and behavior, (8) mind and formal system.

As has been stated repeatedly the questions under discussion by this group concern the criteria for deciding whether something is or is not a computer, and for deciding whether minds are examples of computers. First, I will attempt to remind us all of the role of crucial criteria, thereby laying the groundwork for the methodology of my thinking on the question. Then I will explore the duals mentioned above. Finally, I will attempt to summarize a response to the questions we are discussing.

Popper (e.g., 1962) argued that what distinguishes science from nonscience is the use of a falsificationist strategy. He recognized that one can never PROVE the truth of a conjecture, e.g., there are no black swans, computers are incapable of thinking; but he did argue that one could DISPROVE a conjecture. We could disprove the black swans conjecture by finding a black swan, and we could disprove the computers conjecture by finding or building one capable of thought. There are two very important problems with this view. First, every hypothesis or conjecture has attached to it an implicit ceteris paribus assumption (i.e., all other things being equal). Proving a conjecture to be false requires that we prove the ceteris paribus assumption to be true, that is, that there was no contaminating factor that inadvertently caused the observed results. This is also a conjecture, and we know that we cannot prove its truth, so therefore, observation can neither prove nor disprove a conjecture. Second, say that we found a black bird that seemed to be a swan or found a computer that seemed to think. How do we know that it actually is a swan (although black) or that it actually thinks? These are also conjectures and we know that we cannot prove them to be true. We can apply the Bush Duck Test: if it looks like a swan, and smells like a swan, and tastes like a swan then it is a swan (the TTT for swanness). Although we might agree that this creature appears to be a swan, in fact, we cannot prove it. No matter how many tests we run, the very next test may be inconsistent with the bird being a swan. In fact, like the rest of the conjectures, we cannot prove that this test is appropriate and relevant, so we cannot know for sure that the bird is NOT a swan. The conclusion is that we cannot know for certain whether a conjecture is true or false. Certainty is simply unattainable (see Lakatos & Musgrave, 1970; Moore, 1956). The conclusion for our purposes is that no set of crucial criteria (redundancy intended) can be specified to decide whether a machine is or is not a computer or for deciding whether a mind is or is not a computer.

The argument that there can be no proof of any conjectures, including conjectures of the form: "this is a computer" is very informative regarding my prejudices in this context. I take the notions about the impossibility of proof to be central not only to scientific epistemology, but to everyday epistemology. If our scientific concepts are not so clear-cut and formal, then how, I argue, can we expect our ordinary concepts to be rationally based? The notion that concepts can be represented formally, specifically that thinking involves some kind of proof mechanism seems inconsistent and inappropriate. It was once thought that logic was worth studying not only for its mathematical properties but also because logic is the paradigm for actual human thought. Logic is worth studying, but it is not the paradigm for the psychology of thought (Kahneman & Tversky, e.g., 1982).

The present discussion is lively, in part because contributors are using a number of words in subtly (and not so subtly) different ways. The "groundings" for many of the symbols we use are not shared among contributors (lacking a shared base of grounded symbols, one might argue, makes us collectively dysfunctional, thereby demonstrating the necessity of symbol grounding). Words that are problematic for some of us are used as basic- level concepts by some of the rest. One of these words is consciousness. For some individuals, consciousness is used as a synonym for thinking.

For example, Martin Davis wrote:

Whether a TT-passing computer is in any reasonable sense conscious of what it is doing is not a question we can hope to answer without understanding consciousness.

Pat Hayes wrote:

A human running consciously through rules, no matter how 'mindlessly', is not a computer implementing a program. They differ profoundly, not least for practical purposes.

Michael Dyer wrote:

So there is every indication that consciousness is a folk description for behaviors arising from extremely complex interactions of a very complex subsystems. There are probably a VERY great number of variant forms of consciousness, most of them quite foreign to our own introspective experiences of states of mind. Then we have to decide if "anyone is at home" (and to what extent) in gorillas, in very young children, in our pet dog, in a drugged-out person, etc. etc.

These examples illustrate some of the variety of uses of the concept of consciousness. There seems to be an implicit claim that to think is to be conscious. If this is true, then the question of whether a mind is a computer or whether a computer can be a mind is the question of whether a computer can have consciousness. Notice that I have equated "having a mind" and "thinking." I argue for equating mindedness and thinking, but I argue that consciousness is a red herring. Although Dyer equates consciousness to some complex behavior, in fact, a behavioral-level description of what constitutes consciousness is impossible, because of the large number of behaviors that could be consciousness (or manifestations of it). By a behavioral-level description, I mean one that is couched in terms of movements and physical or kinematic descriptions of them.

Another conflation percolating through the discussion involves grounding and reference. A number of contributors seem to agree that an important characteristic of minds, if not of real computers, is that the symbols in the system must be grounded.

For example, Stevan Harnad wrote:

The sensory grounding hypothesis is simply that eventually the symbolic descriptions can be cashed into terms whose referents can be pick out from their direct sensory projections.

There are several problems with equating grounding with the ability to pick out objects from sensory projections. Among these are (1) the inconsistency in sensory projection that are characteristic hobgoblins of machine vision, and (2) the use of terms that have no referent, but are meaningful. Objects, such as birds, are reasonably easily recognized by humans, despite wide variations in their sensory projections (e.g., in vision, the optical projection of the light reflected from the object on the retina). Designing a computer system that can recognize a bird at any orientation and any condition of flight is extremely difficult. This is an empirical matter, not an introspection, and recognition of other objects can be even more difficult. My point in raising this difficulty is not that computers cannot have vision, but rather to point out that recognizing objects from their sensory impressions is not trivial, and so is unlikely (I think) to be a sound basis for our symbols. Pigeons can be trained to discriminate pictures containing trees in them from pictures that do not contain trees, but might contain flowers, shrubs, plants, people, logs, etc. (e.g., Herrnstein, 1984, 1985). It is easier to train the pigeons to discriminate between such natural categories as trees versus nontrees than it is to train them to discriminate one arbitrary set of pictures from another. One psychologist offered as an explanation of this phenomenon that the pigeon could discriminate tree slides because "they looked like trees," but the other pictures did not. This putative explanation for the birds' performance does not even address the issue because it merely restates the observation without offering any explanation for what constitutes "looking like a tree." My point is not that we need computers to understand the mind, in this case how animals or people recognize objects, rather it is that we cannot assume that biological processes necessarily provide the primitive elements that will allow us to escape from pure computationalism. To rely on picking out objects from among some set of alternative objects itself requires explanation, it is not sufficiently primitive to act as the foundation of the grounding.

Symbol grounding is apparently intended to assure that symbols are not just meaningless marks. As Harnad wrote:

. . . systems are just meaningless squiggles and squoggles unless you project an interpretation . . . onto them.

Many meaningful terms, however, have no referents. These include the function words, and all the abstract nouns (e.g., furniture, truth, beauty), as well as certain other conceptual entities. The most famous conundrum concerning reference and meaning (attributed to Russell, I think) is that involving the Golden Mountain in the sentence, "The Golden Mountain does not exist." If it does not exist then how can it be the subject of the reference? Is the symbol, Golden Mountain, meaningless? A related problem is that two symbols with the same referent must, then have the same meaning and must be substitutable for one another. Although the "evening star" and the "morning star" symbols both refer to Venus, one could believe that the evening star is really a planet without believing that the morning star is really a planet, even though they happen to refer to the same object and both are equally effective at picking out the object in question. Hence, reference, or the ability to pick out an object to correspond to a symbol, is not an adequate basis for assigning a meaning to a word. Additionally, some terms allow us to pick out an object among alternatives, but their semantics is unrelated to the object in question. For example, if I ask you to get me a screwdriver, and you do not know which tool I mean, then the phrase "the yellow one" may allow you to pick out the correct item, but the phrase does not mean anything having to do with screwdrivers. To understand a sentence, one must know more than the referent or meaning of the individual words.

Whatever a computer is, attributing to its symbols properties corresponding to words does not help us to understand what makes those symbols carry any significant weight because words themselves are not solidly enough connected to their meanings or to referents. Consider words such as "tire" that have multiple meanings. One symbol, the string of letters, requires us to pick out two entirely orthogonal referents, one having to do with fatigue and one having to do with wheels. As it turns out, many or even most words in English have multiple meanings or variants of meaning, even if they are not so distinct as "tire." The word "strike," for example, has more than 80 meanings listed in my dictionary. Current investigations in my laboratory suggest that these meanings have family resemblance relations, but do not share a core of essential conceptual features. For example, strike a match, strike a blow, strike a bargain, and strike out for Boston, all use the same symbol, some share the feature that might be labeled (with some trepidation, see below) "hit" (rapid acceleration resulting in a collision), but two of them seem completely independent of hitting. The context is important in determining which meaning or which referent to assign to that symbol. The symbol is only grounded in context and cannot be treated simply as a formal object whose use is governed solely by its shape. There may be some internal symbol that maps onto this surface symbol that is not ambiguous, but positing such an internal symbol seems to be an ad hoc adjustment to a hypothesis that relies on external relations to specify the intension of the symbol (e.g., acting appropriately as in the TTT).

Harnad wrote:

I don’t know what the ground-level elementary symbols will turn out to be, I’m just betting they exist -- otherwise it’s all hanging by a skyhook. Nor do I know the Golden Mountain conundrum, but I do know the putative "vanishing intersections" problem, according to which my approach to grounding is hopeless because not even sensory categories (not to mention abstract categories) HAVE any invariants at all: My reply is that this is not an apriori matter but an empirical one, and no one has yet tried to see whether bottom-up sensory grounding of a TTT-scale robot is possible. They’ve just consulted their own (and their subjects’) introspections on the matter. I would say that our own success in categorization is some inductive ground for believing that our inputs are not too underdetermined to provide an invariant basis for that success, given a sufficiently powerful category learning mechanism.

As an aside, it seems to me that investigations of how people actually use words, as opposed to how a robot might use them is not introspection, but rather good empirical research. People really do use words in a variety of ways, they do not merely think that they do. Our words in isolation may be too underdetermined to provide an invariant basis for success, but words in context are obviously understood (much of the time). Further, our success at categorizing is indeed an inductive ground for believing that we categorize, but it does not imply that categorization occurs on the basis of invariant features in the sensory projections. There are other approaches to concept representation and object recognition. Finally, machine vision projects do indeed attempt to recognize objects based on their sensory projections and many of them find the job easier when other information is also included.

For many people the word "symbol" denotes a discrete item that can be used systematically in a formal system. Several contributors, for example, seem to accept the dichotomy between symbolic systems and analog systems.

Harnad wrote:

Retinal transduction and the analog transformations that follow from it are computer simulable, they are equivalent to computation, but they are not eo ipso computational.

In contrast, Mclennan wrote (I think correctly):

My second point is that the nature of computation can be illuminated by considering analog computation, because analog computation does away with discrete symbols, yet still has interpretable states obeying dynamical laws. Notice also that analog computation can be formal in exactly the same way as digital computation. An (abstract) analog program is just a set of differential equations; it can be implemented by a variety of physical devices, electronic, optical, fluidic, mechanical, etc.

Analog relations are still symbolic in that the state of the analog system represents the state of the object being represented, but the relations are more continuous and less arbitrary than those of a strict, discrete, formal system. There is no reason to believe that human cognition is based on discrete symbols and there is good evidence that human cognition is based on more continuous representations that are ad hoc cast into discrete categories (e.g., Barsalou, 1983, 1987; Labov, 1973). For example, human performance of many kinds suggests that people confuse items that are represented similarly more often and more completely than items that are represented less similarly. The relations between items as remembered is not arbitrary and formal, but is related to the items as perceived and to the context in which they were perceived. This is not introspection, this is a summary of behavior.

Harnad has proposed what he calls the Total Turing Test (TTT) as the crucial experiment to decide whether a computer can have a mind. His claim is that a necessary feature for a mind is the ability to interact with the world both perceptually and behaviorally. Therefore, no artifact can have a mind unless it can behave in ways that are indistinguishable from the way a human would behave in the same situation. I hope that it is clear already that such a test cannot be truly definitive, if for no other reasons than one cannot prove that there are no differences and because human behavior is not regular enough to allow any finite experiment to give an adequate test of the hypothesis, and finally, we do not and, I believe, cannot, have a definitive catalog of situated human behavior. The best we can hope for from a robot in a TTTest is that it behaves in a more or less human like manner.

Harnad wrote:

Sensory grounding cannot be investigated by armchair introspection on word meanings; it will only be understood through empirical attempts to design grounded systems.

The position that the only way to understand symbol grounding is to build robots is an interesting one. Building (or simulating) robots is very useful for investigating a wide range of theories and hypotheses, nevertheless, despite my support for robotics, it is not the only way to be empirical about symbol grounding. Data concerning symbol grounding come from many sources, not just from attempts to build interactive simulations. The simulations themselves cannot be attempted willy-nilly but must be based on one or more emerging theories of fundamental intelligence.

Summary of the comments so far: Proof of scientific conjecture is impossible. No set of observations can ever prove scientific conjecture or theory to be true. As a result, formal systems do not provide a solid epistemological basis for scientific nor for everyday concepts. Symbol grounding has something to do with establishing the semantics or meaning of the symbols, but reference is a weak method for establishing such semantics. Symbols need not be discrete, but can be continuous, based on analog relationships. Finally, formal systems are not a good paradigm for capturing biological intelligence.

Tearing down someone else's carefully constructed edifice is easy. Building a substitute that will shelter us from the cold of ignorance is considerably more difficult. I offer the following suggestions for a substitute approach to the problem of understanding the relation between minds and computers. The underlying assumption is that a computer can convincingly have a mind if it can function more or less along the lines that biological minds employ. This requires that we understand the biological minds we are trying to match as well as understand the computational minds we might try to develop.

Harnad has warned us, to some extent correctly, that any enterprise of mind construction will be fruitless unless the system's symbols are grounded. At a minimum, grounded symbols require that the system behave systematically relative to some environment. I have argued that the methods that have been suggested so far for symbol grounding are inadequate and to some extent inappropriate. The paradigms that many of us have adopted for understanding the mind and computer are inappropriate. There are two parts to my suggested approach. The first is methodological, and the second is representational.

A number of investigators in recent years have pursued an approach to understanding intelligence that some of us have come to call the biomimetic approach. Rather than focus on modeling performance of those tasks that characterize so-called higher human intelligence, such as planning, problem solving, scientific creativity, and the like, this approach focuses on the complementary comparative approach of modeling whole, albeit simple, organisms in a real environment, performing real biological tasks. The goal of the approach is to develop coherent incremental models out of functionally complete components. The more common approach has been successful at modeling those tasks that humans find difficult and perform slowly (such as expert chess playing), but which can be described according to specified rules. The tasks tend to be rather small portions of the whole of human performance, and to operate on the basis of a limited range of inputs that are often a restricted set of simple assertions abstracted from real data (e.g., photos of a scene) by human investigators. Such verbal-like systems have not been as successful in modeling tasks that humans find easy and automatic, such as recognizing the face of a friend.

The biomimetic approach seeks to begin with the development of simple systems that are capable of surviving and operating in a real environment. The goal is then to gradually scale up the models to wider ranges of environments and capabilities, at each level producing systems that are capable of dealing with the essential biological tasks in their environment (e.g., Brooks, 1991). Similarly, understanding the whole of human behavior may simply be too complex to be understood without more understanding of the underlying basic cognitive processes that may be more accessible in animals. Success in explaining human performance may depend a great deal on understanding the operation of fairly basic processes because these processes constrain and support the operations of those so-called higher cognitive functions. The use of animals and their behavior as systems to be modeled helps us to see alternative forms of representation that are overshadowed by our own linguistic capacities and our own introspective self-familiarity. We are less ready to believe, for example, that animals employ formal rule systems, than we are to believe that humans employ formal rules. Nevertheless, both display some level biological intelligence.

The representational part of the approach I suggest is to abandon formal discrete symbols as the basis for representation and abandon reference as the means for grounding those symbols. A major reason for wanting to use formal systems is that the processes by which truth is transmitted and preserved are well understood in formal systems. Monotonic logic guarantees that truth is transmitted from a set of premises (grounded symbols) to a set of conclusions. Abandonment of monotonic logic means that in principle any conclusions are possible. Nevertheless, monotonic logic seems a very poor choice for a paradigm of human thought. Just as scientific epistemology could recognize that there are no guarantees of truth in scientific concepts without dissolving into irrationality, our models of human thought can tolerate nonmonotonic logic. My argument is that thought does not operate as a formal system operating with discrete symbols and proof of syllogisms, rather it operates as a more or less continuous system operating on the basis of constraint satisfaction. I will illustrate what I mean by reference to word understanding, but similar mechanisms can be applied to many forms of behavior (Chemtob, et al., 1989; Roitblat, 1990).

I argue that concepts and words are represented in a very high- dimensional space in which the dimensions are semantic. Any given word is represented as a cloud of points in this space. For example, the word "strike" is represented along many dimensions corresponding to the different semantic aspects of the word. Because of the discontinuity from one use of strike to the next, some representations of the word have positions along a "hit" dimension that indicate a forceful blow (as in "he struck and killed the pedestrian") and some have locations that indicate no blow (as in "they struck a bargain"). (Note that the labels that I can place on these dimensions are only a poor description of the semantic dimension, because the labels themselves have multiple and multidimensional meanings and can only approximate the true underlying dimension.) Obviously these representations of the word strike are incompatible with one another and people are unlikely to be able to employ both meanings simultaneously (there is evidence on this topic). When a person recognizes the word strike, I argue, all of these dimensions are activated simultaneously, but because they cannot all remain active, a projection of the representation is performed from the very high dimensional space onto a space of lower dimensionality. The dimensions that are selected for the projection are those that are most consistent with the context. Hence, word understanding is an iterative constraint satisfaction process in which the meaning of the word that best fits the context is the one that remains active. What appears to be a circularity problem turns out to be an iterative problem. The meaning of the words constrain the meaning of the sentence and the meaning of the sentence constrains the meanings of the words. The words are not firmly grounded in isolation, but neither are they merely hung from skyhooks.

An iterative continuous constraint satisfaction system is not guaranteed to solve the problems of symbol grounding, computation, and artificial intelligence. No system can offer such a guarantee. Nevertheless, it appears to offer an alternative path toward solution of such problems.

References Barsalou, L. W. (1983). Ad hoc categories. Memory and Cognition, 11, 211- 227. Barsalou, L. W. (1987). The instability of graded structure: implications for the nature of concepts. In U. Neisser (Ed.), Concepts and conceptual development: Ecological and intellectual factors in categorization (pp. 101-140). New York: Cambridge University Press.

Brooks, R. A. 1991. Intelligence without representation. Artificial Intelligence, 47:139-159.

Chemtob, C., H. L. Roitblat, R. S. Hamada, J. G. Carlson, and G. T. Twentyman 1991. A Cognitive Action Theory of Post- Traumatic Stress Disorder. Journal of Anxiety Disorders, 2:253-275.

Herrnstein, R. J. 1984. Objects, categories, and discriminative stimuli. In H. L. Roitblat, T. G. Bever, & H. S. Terrace (Eds.), Animal Cognition (pp 233-262). Hillsdale, NJ: Lawrence Erlbaum Associates.

Herrnstein, R. J. (1985) Riddles of natural categorization. Philosophical Transactions of the Royal Society (London) B 308: 129-44

Labov, W. (1973) The boundaries of words and their meanings. In C.-J. N. Bailey & R., W. Shuy (Eds.), New ways of analyzing variations in English. Washington, DC: Georgetown University Press.

Lakatos, I. & Musgrave, A. (1970) Criticism and the growth of knowledge. Cambridge: Cambridge University Press.

Kahneman, D. & Tversky, A. (1982) On the study of statistical intuitions. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgements under uncertainty: Heuristics and biases (pp. 493-508). Cambridge, Cambridge University Press.

Moore, E. F. (1956) Gedanken-experiments on sequential machines. In C. E. Shannon & J. McCarthy (Eds.), Automata studies (pp. 129-153). Princeton: Princeton University Press.

Popper, K. P. (1962) Conjectures and refutations. New York: Harper and Row.

Roitblat, H. L. (1988) A cognitive action theory of learning. In J. Delacour and J. C. S. Levy (eds.), Systems with learning memory abilities. New York:Elsevier. pp. 13-26.

Roitblat, H. L. (1990) Cognitive action theory as a control architecture. In S. Wilson and J. A. Meyer (Eds.), Simulation of Adaptive Behavior from Animals to Animats. MIT Press, Cambridge, Mass., pp. 444-450.

Date: Wed, 20 May 92 22:18:41 EDT From: "Stevan Harnad"

Below are 5 more responses to the question about publishing the "What is Computation" Symposium, for a total of 19 votes cast out of total of 25 contributors (at the time of the vote):

Publication: For: 18 // Against: 1

Interactive Symposium (IS) vs. Position Papers (PP): Either or Combination: 11 - Prefer IS: 5 - Prefer PP: 2

Not yet heard from (6):

(20) Ross Buck (21) Ronald Chrisley (22) Gary Hatfield (23) Joe Lammens (24) John Searle (25) Tim Smithers (and any new contributors since the vote)

I do wish to remind contibutors that we are still in the interactive symposium phase, no matter what the outcome, so please do NOT send lengthy position papers yet: Keep it interactive and about the length most contributions have been thoughout the discussion.

There is also some question about how to partition the two themes ("What is Computation?" and "Is Cognition Computation?") in the published version, and whether to include the second theme at all. (I personally think it would be hard to eradicate all traces of the second question, which is in any case in the back of most of our minds in all of this). -- Stevan

----

(15) Date: Wed, 13 May 92 16:16:04 PDT From: dambrosi@research.CS.ORST.EDU (Bruce Dambrosio)

Stevan -

My short comment hardly qualifies me to render an opinion, I’m neutral. I do hope, however, that the precedent of publication doesn’t change the flow of future discussion. thanks - Bruce

----

(16) Date: Thu, 14 May 92 10:31:07 EDT From: "John M. Carroll"

stevan since you registered me as a voter, i’ll vote. i’m just an applied ontologist in the audience of this symposium, but i’ve found it interesting (though i do agree with mcdermott -13 may- as to its wandering itinerary). publishing it as ’position papers’ would seem to me to factor out some of the unique value of interactive debate (though squeezing out some of the asynchrony and redundancy that come along with e-mail is probably a good idea). bravo and onward. jack

----

(17) Date: Fri, 15 May 92 18:31:54 -0400 From: mclennan@cs.utk.edu

Stevan,

Publishing the "What is Computing" dialogue is fine with me. I do think it will need some smoothing to avoid being too repetitious, but it should be possible to do.

Bruce MacLennan Department of Computer Science The University of Tennessee Knoxville, TN 37996-1301

----

(18) From: sjuphil!tmoody@uu.psi.com (T. Moody) Date: Mon, 18 May 92 13:36:32 EDT

Stevan,

Publishing this discussion in some edited form is an excellent idea. I regret that I have not been more active, but I have gotten a great deal out of reading thing, and I am sure that others would, too.

-- Todd Moody

----

(19) Date: Tue, 19 May 1992 21:22:37 From: Pat Hayes

Stevan - yes, Im here , but I am out of touch with email for days at a time. This will continue for about another two weeks. Sorry. I approve of the project and will try to get back into it ina few days.

Pat

Date: Wed, 20 May 92 22:52:16 EDT From: "Stevan Harnad"

Date: Wed, 20 May 1992 17:03:14 -0400 (EDT) From: Franklin Boyle

Stevan Harnad writes (in response to Dave Chalmers):

>
\>sh> I think the word "structure" is equivocal here. A computer simulation
>
\>sh> of the solar system may have the right causal "structure" in that the
>
\>sh> symbols that are interpretable as having mass rulefully yield
>
\>sh> symbols that are interpretable as gravitational attraction and
>
\>sh> motion. But there’s no mass, gravity or motion in there, and
>
\>sh> that’s what’s needed for REAL causality. In fact, the real
>
\>sh> causality in the computer is quite local, having to do only
>
\>sh> with the physics of the implementation (which is irrelevant to
>
\>sh> the computation, according to functionalism). So when you
>
\>sh> speak equivocally about a shared "causal structure," or about
>
\>sh> computational structure’s being a "variety of causal structure," I
>
\>sh> think all you mean is that the syntax is interpretable AS IF
>
\>sh> it were the same causal structure as the one being modelled
>

>sh> computationally. In other words, it’s just more ungrounded,
>
>sh> extrinsic semantics.

Well said. To elaborate, much of the physics of the system depends on the causal behavior of electric charge. But the combinations of ’high’ and ’low’ voltage values that instantiate symbols *control* the computationally relevant physical changes; e.g., a voltage change which opens a data path from the starting symbol of a subroutine to a register in the CPU. The symbols cause these changes as a result of their structures via pattern matching.

Though each individual voltage value that is part of a symbol’s physical instantiation causes change according to physical laws, the voltage combinations are able to cause changes AS THOSE STRUCTURES because of the circuit architectures of electronic devices, such as comparators. The structures of these devices enable all the individual, nomologically determined electrical changes taken together to result in an overall change (e.g., a simple change from a high to low voltage) which reflects the similarity between the *arrangement* of voltages that constitutes the structure of the instantiated symbol and that of its matcher. This latter, overall change is not described by physical laws. Rather, the regularities such changes generate are described by rules (because symbols cause change according to their extended structures) and, hence, underlie the computational behavior of digital computers.

In most physical systems (enzyme catalysis in cells, for example, and, of course, digital computers are two exceptions), there is no structure fitting, only nomologically determined changes in the measured attributes of interacting objects (e.g., momentum). This is what happens with planets in their orbits as well as in analog computers. These systems behave according to REAL causality, as you put it, precisely because the changes are nomologically determined from the measured attributes of the interacting objects themselves. In contrast, computer simulations are simulations because measured-attribute changes are not the result of the measured attibutes of the interacting objects (in this case symbols), but, rather, their extended structures, and, furthermore, because such changes do not affect the physical aspects of the interacting objects themselves, as they do in the case of planetary motion. That is, symbols control the manipulation of other symbols (e.g., moving them around in memory) by controlling changes in measured attributes (voltages) of particular circuits, but not of each other.

On the other hand, if symbols do not cause changes in those symbols through pattern matching, then they are not affecting system behavior by virtue of their forms, and, so, the system would simply not be a computer in that case. Thus, planets in orbits are not computers.

If we describe digital computers as we would any other physical system, that is, in terms of physical state descriptions, it would relegate the structures of the symbols piecemeal to boundary conditions (like physical laws, these boundary conditions are associations between measured attributes -- state variables and structural quantities). Such descriptions, therefore, would miss the fact that certain voltage changes are the result of structure fitting, and, hence, would not capture the computational aspects of digital computers because they would not capture the causality due to the symbols.

Searle’s statement, "syntax is not intrinsic to physics", summarizes quite nicely the fact that physical state descriptions, which are the standard way of describing the world because they are about measured attributes of physical objects which physical laws associate, do not capture the structural causality of syntactic structures; that is, structures which are causal via pattern matching. In other words, the physical behavior of computers can be described by a set of integro-differential equations and constraint equations without ever having to account for the causality of extended structure in any explicit way.

>
>sh> I think I can safely say all this and still claim (as I do) that
>
>sh> I accept the Church/Turing Thesis that computation can simulate
>
>sh> anything, just as natural language can describe anything.
>
>sh> We just mustn't confuse the simulation/description with the real
>
>sh> thing, no matter how Turing-Equivalent they might be. So if we
>
>sh> would never mix up an object with a sentence describing it, why
>
>sh> should we mix up an object with a computer simulating it?

Great. Actually I would have changed the first sentence to: "...that computation can simulate anything that can be described...", precisely because digital computers enable descriptions (usually in the form of patterns which can be structurally matched by rule antecedents) to be causal.

Franklin Boyle

Date: Tue, 19 May 92 20:57:02 EST From: David Chalmers

There are too many deep issues here to treat them in any anywhere near the depth they deserve, but here goes. I'll start with computation and move up upwards through the rarefied heights of cognition, semantics, and qualia.

(1) WHAT IS COMPUTATION?

>
>sh> I certainly couldn't agree with you on computation without dissociating
>
>sh> myself from this part of your view. But let me, upon reflection, add
>
>sh> that I'm not so sure your criterion for computation does the job (of
>
>sh> distinguishing computation/computers from their complement) after all
>
>sh> (although I continue to share your view that they CAN be distinguished,
>
>sh> somehow): I don't see how your definition rules out any analog system
>

>sh> at all (i.e., any physical system). Is a planetary system a computer
>
>sh> implementing the laws of motion? Is every moving object implementing a
>
>sh> calculus-of-variational computation? The requisite transition-preserving
>
>sh> mapping from symbols to states is there (Newton's laws plus boundary
>
>sh> conditions). The state transitions are continuous, of course, but you
>
>sh> didn't specify that the states had to be discrete (do they?).

A planetary system is not a computer, because it's not universal. I should also note that computation requires counterfactual sensitivity to various different possible inputs, and it's not clear what will count as an "input" to the solar system. But apart from that worry, there's no problem with saying that the solar system is implementing any number of specific computations, e.g. the trivial 1-state FSA as well as a lot of cyclic n-state FSAs. It's probably not implementing a calculus-of-variations computation, as such a computation would require a particular kind of state-transitional structure that this system does not embody. (There may be some sense in which the system is "I/O" equivalent to such a computation, but computational equivalence requires more than I/O equivalence, of course.)

Remember that it's not a problem for my view that every system implements some computation or other. What matters is that every system does not implement *every* computation.

As for continuity or discreteness, that depends on the computational formalism that one uses. Certainly all of the usual formalisms use discrete states. Of course, a continuous physical system (like the planetary system) can implement a discrete computation: we just have to chop up its states in the right way (e.g. divide an orbit into 4 discrete quadrants).

>
>sh> And what about syntax and implementation-independence, which are surely
>
>sh> essential properties of computation? If the real solar system and a
>
>sh> computer simulation of it are both implementations of the same
>
>sh> computation, the "supervenient" property they share is certainly none
>
>sh> of the following: motion, mass, gravity... -- all the relevant
>
>sh> properties for being a real solar system. The only thing they seem to
>
>sh> share is syntax that is INTERPRETABLE as motion, mass, gravity, etc.
>
>sh> The crucial difference continues to be that the interpretation of being
>
>sh> a solar system with all those properties is intrinsic to the real solar
>

>sh> system "computer" and merely extrinsic to the symbolic one. That does
>
>sh> not bode well for more ambitious forms of "supervenience." (Besides, I
>
>sh> don't believe the planets are doing syntax.)

This certainly isn't an objection to my construal of computation. It's *computation* that's implementation-independent, not, e.g., solar-system-hood. The fact that the solar system might be implementing a computation is not affected by the fact that other implementations of that computation aren't solar systems.

Most properties in the world don't supervene on computational structure, as they don't even supervene on causal organization. To be a process of digestion, for instance, more than a certain causal organization is required: what's also needed is a specific kind of physio-chemical makeup. This physio-chemical makeup is *conceptually constitutive* (in part) of something's being digestion. Similarly for solar systems. It's conceptually constitutive of solar-system-hood that a system have a certain geometric shape, a certain chemical makeup, a certain size, and so on, and these physical properties are not determined by abstract causal organization. Take a system that shares abstract causal organization with a solar system -- the Bohr atom, say, or a boy swinging a bucket around his head -- then it's still not a solar system, because it lacks those extra properties that are constitutive of solar-system-hood. So no one would dream of being a computationalist about solar-system-hood, or about digestion.

The strong-AI hypothesis is that unlike these properties, *cognition* is a property that supervenes on abstract causal organization. This may or may not be obvious at first glance, but note that unlike digestion and solar-system-hood, it's not ruled out at first glance: there doesn't seem to be any physical property independent of causal organization that's conceptually constitutive of cognition.

In general, computational simulation will succeed at most in duplicating those properties that supervene on causal structure. We can argue all day about whether cognition is such a property, but the important point here is that pointing to properties that don't supervene on causal structure is no objection to my construal of computation.

>
>sh> I think the word "structure" is equivocal here. A computer simulation
>
>sh> of the solar system may have the right causal "structure" in that the
>
>sh> the symbols that are interpretable as having mass rulefully yield
>
>sh> symbols that are interpretable as gravitational attraction and motion.
>
>sh> But there's no mass, gravity or motion in there, and that's what's
>
>sh> needed for REAL causality. In fact, the real causality in the computer
>
>sh> is quite local, having to do only with the physics of the implementation
>
>sh> (which is irrelevant to the computation, according to functionalism).

>
>sh> So when you speak equivocally about a shared "causal structure," or
>
>sh> about computational structure's being a "variety of causal structure," I
>
>sh> think all you mean is that the syntax is interpretable AS IF it were
>
>sh> the same causal structure as the one being modelled computationally. In
>
>sh> other words, it's just more, ungrounded, extrinsic semantics.

Not at all. I mean that every implementation of a given computation has a *real* *causal* *structure*, and in fact that there's a certain causal structure that every implementation of a given computation shares. That's precisely what the definition of implementation guarantees. When a given 2-state FSA is implemented on my computer, for instance, there are real physical state-types in the implementation such that being in state A causes a transition into state B, and vice versa. When a neuron-by-neuron simulation of the brain is implemented on my computer, there are real physical states (registers, or memory locations, or whatever) in the implementation corresponding to the state of each neuron, and these states interact with each other in a causal pattern isomorphic to a pattern of interaction among the neurons.

To clarify, by "causal structure" I mean, roughly, *organizational* properties of a system: i.e., the patterns of interactions between various states, without taking into account what those states actually are. For instance an atom, at least according to the Bohr model, might share some causal structure with the solar system, but it differs in many properties that aren't organizational properties, such as size, mass, and intrinsic physical structure.

This has to be kept quite separate from questions about semantics. I haven't yet even mentioned any possible associated "semantics" of the computation. And again, I wouldn't dream of claiming that a simulation of the solar system has the same *gravitational* properties as the solar system, or that a simulation of the brain has the same *biochemical* properties as the brain, and I don't know why you think this is implied by my position. Gravitation and biochemistry don't supervene solely on causal structure, obviously.

(2) COMPUTATION AND COGNITION

We now pass briefly to the question of whether *cognition* might be a property that supervenes on causal structure, and on computational structure in particular.

>
>sh> There is a straw man being constructed here. Not only do all
>
>sh> Functionalists agree that mental states depend on causal structure, but
>
>sh> presumably most nonfunctionalist materialists do too (neurophysical
>
>sh> identity theorists, for example, just think the requisite causal
>
>sh> structure includes all the causal powers of -- and is hence unique to

>
>sh> -- the biological brain).

Well, no, at least not the way I'm using "causal structure". Given any specification of the causal structure of the brain -- even all the way down to atoms, or whatever -- then that causal structure could in principle be implemented in a different medium, such as silicon. We'd just have to set it up so that our little bits of silicon are interacting with each other according to the same patterns as the neurons, or the atoms or whatever, were interacting with each other. (Of course the silicon model might be a lot bigger than the brain, and it might have a lot of *extra* causal structure that the brain doesn't have, but that's not a problem.) Now a neurophysiological identity theorist would certainly say that the silicon system wouldn't have the same mental states. So the way I'm using the term (and I think this is standard usage), a neurophysiological identity theorist would not agree that mental states supervene on causal structure.

Perhaps you don't agree that mental states depend solely on causal structure either, because you seem to assign an essential role to I/O transducers, and presumably it makes a difference just what kinds of physical things -- heat, light, or whatever -- are being transduced. Whereas a strict functionalist like myself would hold that at least when it comes to fixing phenomenal mental states, the specific physical nature of what's being transduced is irrelevant. On this view, a system that merely reproduced the causal organization of the transduction in a different medium would have the same phenomenal properties.

As I said in the last note, even if one accepts (a) that computational structure fixes causal structure (which follows from my construal of implementation), and (b) that causal structure fixes mental structure, there still arises the question of whether computational structure can fix the right *kinds* of causal structure that are responsible for mentality. I think that it can: we just have to capture the causal structure of the brain, say, at a fine enough level of description, and describe that causal structure in an appropriate computational language -- as a finite state automaton, for instance, though preferably as an FSA with combinatorially structured states. Then every implementation of that FSA will share that causal structure. Some people might hold that *no* finite level of description can capture everything that's going on, due e.g. to the potential infinite precision in continuous systems, but I think that the presence of background noise in biological systems suggests that nothing essential to cognition can ride on that infinite precision.

Before passing on to the next topic, I should note that I don't think that "Is cognition computation?" is quite the right question to ask. The right question, rather, is "Is computation sufficient for cognition?" An advocate of strong AI might reasonably hold that cognition in the brain is not itself computation, but that computation is nevertheless capable of reproducing the relevant properties (e.g. causal structure) on which cognition depends. This becomes particularly clear when we move to specific computational formalisms, such as Turing machines. I certainly don't think that the brain is a Turing machine, but I think that nevertheless Turing machine computations are capable of cognition. It's a subtle point, but too often advocates of AI are saddled with unnecessary claims such as "the brain is a computer", or "the mind is a program".

(3) COMPUTATION AND SEMANTICS

>
>sh> "Syntactic" means based only on manipulating physical symbol tokens
>

>sh> (e.g., squiggle, squoggle) whose shape is arbitrary in relation to what
>
>sh> they can be interpreted as meaning. I am sure one can make
>
>sh> squiggle-squoggle systems, with arbitrary formal rules for manipulating
>
>sh> the squiggles and squoggles -- like Hesse's "Glass Bead Game" but even
>
>sh> more absurd, because completely meaningless, hence uninterpretable in
>
>sh> any systematic way -- and one could perhaps even call these
>
>sh> "computations" (although I would call them trivial computations). But I
>
>sh> have assumed that whatever it turns out to be, surely one of the
>
>sh> essential features of nontrivial computations will be that they can
>
>sh> bear the systematic weight of a semantic interpretation (and that
>
>sh> finding an interpretation for a nontrivial symbol system will be
>
>sh> crytographically nontrivial, perhaps even NP-complete).

The question is only whether semantic content is itself *constitutive* of something's being a computation. To that question, the answer seems obviously to be no. Construct an arbitrary large Turing machine by throwing together quadruples randomly. It's most unlikely that there will even *be* a nontrivial semantic interpretation for this. Construct a Pascal program by making random decisions consistent with the BNF specification of the language. Almost certainly, this program won't be interpretable as being about anything at all. Nevertheless, it's still a *program*, and an implementation of it is still *computation*, at least according to standard usage, which I think is the right usage. It's probably not a very interesting computation, but it's computation.

Most interesting computations will probably turn out to have some kind of semantic interpretation -- otherwise why would we bother with them? (Actually, some interesting computations might not, e.g. those computations invoked in solving the "Busy Beaver" problem for Turing machines. These computations are interesting, but the interest appears to lie entirely in their syntax. Similarly, many cellular automata computations, like Conway's game of life, are interesting primarily for their syntactic form.) But the notion that lies at the foundation of the computationalist view about cognition is not "interesting computation", it's "computation" straight. Making some sense of the notion of "interesting computation" is an interesting question in its own right, but it's independent of Searle's original question about what makes something a computation.

(4) QUALIA AND SEMANTICS.

Now we move away from computation to the nitty-gritty philosophical questions about different kinds of mental properties. Unlike the issues about computation and implementation (which are surprisingly underrepresented in the literature), these issues already have a vast philosophical literature devoted to them. What I'll have to say here won't be particularly original, for the most part.

It’s also more philosophically technical than the earlier parts, so some readers might want to drop out now (if they’ve made it this far, which is unlikely).

>
>sh> At some point (mediated by Brentano, Frege and others), the mind/body
>
>sh> problem somehow seems to have split into two: The problem of "qualia"
>
>sh> (subjective, experiential, mental states) and the problem of
>
>sh> "intentionality" (semantics, "aboutness"), each treated as if it were
>
>sh> an independent problem. I reject this bifurcation completely. I
>
>sh> believe there is only one mind/body problem, and the only thing that
>
>sh> makes mental states be intrinsically about anything at all is the fact
>
>sh> that they have experiential qualities.

To set out the lay of the land, I agree that there’s only one mind-body Problem worthy of a capital P, and that’s the problem of qualia. That’s not to say that qualia are the only kind of mental states: as I outlined in my last post, there are also "psychological states", those characterized by their role in the production of behaviour rather than by their phenomenal feel. However, there’s no more a mind-body Problem about these than there is a "life-body Problem", for instance. The very fact of a system’s being alive is a fact about it’s incorporating the right kinds of mechanism and producing the right kind of behaviour (where the key behaviour and mechanisms are adaptation, reproduction, and metabolism, more or less). There’s no "further fact" that needs explaining. The same goes for psychological states. What’s special about qualia, and makes them seem unlike almost everything else in the world, is that there seems to be a further fact in need of explanation, even after one has told the full story about the mechanisms and so on.

Where I differ from you is in assimilating intentional states like beliefs to the class of psychological states, rather than to the class of phenomenal states. *All there is* to the fact of a system believing that P is that it has the right kind of causal economy, with mechanisms that tend to produce P-appropriate behaviour in the right sort of ways, and that are causally related to the subject matter of P in the right sort of way. The possession of semantic content isn’t a further fact over and above these mechanisms: it *conceptually supervenes* on the existence of those mechanisms, to use the philosophical parlance.

>
>sh> If there were nothing it was like (subjectively) to have beliefs and
>
>sh> desires, there would be no difference between beliefs and desires that
>
>sh> were just systematically interpretable AS IF they were about X
>
>sh> (extrinsic semantics) and beliefs and desires that were REALLY about X
>

>sh> (intrinsic semantics).

One might parody this argument by saying:

If there were nothing it was like (subjectively) to be alive, there would be no difference between systems that were just systematically interpretable AS IF they were alive (extrinsic life) and systems that were REALLY alive (intrinsic life).

Obviously, any system that is functioning in the right way is not just "as if" alive, it's really alive, qualia or no qualia. The same goes for belief. Maybe this means that there's not much difference between "as if" believing and "real" believing, but why should that bother us? We don't worry about a difference between "as if" tables and "real" tables, after all.

(There does remain one "as if" vs. "real" distinction, which is that a system might *behave* as if it believes that P without believing that P (actors do this, for instance, and Block's fabled Humongous Lookup Table might do it even better). But this problem can be handled without invoking qualia: functionalism requires that to determine the facts about a belief, one must invoke not only the facts about behaviour but the facts about patterns of internal causation. The patterns of causation in actors and lookup tables don't qualify. Spelling out the right criteria on internal causation is a long, intricate story, but qualia don't need to be invoked anywhere.)

>
>sh> There are qualia, however, as we all know. So even with a grounded
>
>sh> TTT-capable robot, we can still ask whether there is anybody home in
>
>sh> there, whether there is any haver of the beliefs and desires, to whom
>
>sh> they are intrinsically [i.e., subjectively] meaningful and REALLY about
>
>sh> what they are interpretable as being about. And we can still be dead
>
>sh> wrong in our inference that there is somebody home in there -- in which
>
>sh> case the robot's semantics, for all their causal groundedness, would in
>
>sh> reality be no more intrinsic than those of an ungrounded book or
>
>sh> computer.

Qualia or no qualia, beliefs are still "intrinsic" (modulo questions about narrow and wide content), in just the same way that life is intrinsic. It's just that they're not *phenomenal*.

The fundamental problem with making qualia essential to semantic content is that qualia seem to be *the wrong kind of thing* to determine that content (except perhaps for certain kinds of perceptual content). As I said earlier, my belief about Joan of Arc may have some associated (though hard to pin down) qualia, but it's very difficult to see how those qualia are *constitutive* of the semantic content of the belief. How could the *feel* of the belief possibly make it any more about Joan of Arc than it would have been otherwise?

Your position, I take it, is roughly that: "as if" semantic content *plus* qualia *equals* "real" semantic content. My position is that qualia seem to contribute almost nothing to fixing the semantic content of most beliefs, except perhaps for certain perceptual beliefs. So whatever it is that is constitutive of "real" semantic content, qualia don't play much of a role. This may mean that there won't be much of a "real"/"as if" distinction to worry about (modulo the considerations about behavioural equivalence), but that's life.

> dc> Your position seems to be, on the contrary, that qualia are
> dc> determinative of semantic content. Take Joe, sitting there with some
> dc> beliefs about Joan of Arc. Then a hypothetical system (which is at
> dc> least a conceptual possibility, on your view and mine) that's
> dc> physically identical to Joe but lacks qualia, doesn't believe anything
> dc> about Joan of Arc at all. I suggest that this seems wrong. What can
> dc> qualia possibly add to Joe's belief to make them any more about Joan
> dc> than they would have been otherwise? Qualia are very nice things, and
> dc> very important to our mental life, but they're only a matter of *feel*
> dc> -- how does the raw feel of Joe's belief somehow endow it with semantic
> dc> content?
>
>
>sh> But Dave, how could anyone except a dualist accept your hypothetical
>
>sh> possibility, which simply amounts to the hypothetical possibility that
>
>sh> dualism is valid (i.e., that neither functional equivalence nor even
>
>sh> physical identity can capture mental states!)?

Well, I'm only saying that this is a *conceptual* possibility, which surely it is on your view and mine, not an empirical possibility. I have little doubt that as an empirical fact, any system physically identical to me will have the same qualia. But it's entirely coherent to *imagine* a system physically identical to me but lacking qualia. Indeed, if it wasn't for first-person knowledge of qualia, one would never suspect that such a brain-structure would have qualia at all! (Note that someone (like Lewis or Armstrong, or Dennett on one of his less eliminativist days) who holds that all there is to the *concept* of qualia is the notion of a state that plays a certain causal role won't accept this. But this view simply seems to legislate the problem of qualia into something else entirely.) Qualia are individuated by their phenomenal feel, which seems to be conceptually independent of any physical properties.

So far this view doesn't immediately imply dualism. At least, many people who take qualia seriously accept this conceptual possibility, but still think that ontologically, qualia aren't anything over and above the physical. Personally, I find this view untenable, and think that the conceptual possibility of absent or inverted qualia must imply at least a limited kind of ontological dualism (so-called property dualism), as it implies that there are contingent facts about the world over and above the physical facts. But let's not go into that, for now.

> dc> I suggest that there is some kind of conceptual confusion going on
> dc> here, and that phenomenal and semantic properties ought to be kept
> dc> separate. Intentional states ought to be assimilated to the class of
> dc> psychological properties, with their semantic content conceptually
> dc> dependent on their role in our causal economy, and on their causal
> dc> relations to entities in the external world.
>
>
>sh> Apart from real TTT interactions, I don’t even know what this passage
>
>sh> means: what does "assimilated to the class of psychological properties
>
>sh> with their semantic content conceptually dependent on their role in our
>
>sh> causal economy" mean? "[T]heir causal relations to entities in the
>
>sh> external world" I can understand, but to me that just spells TTT.

I’m not sure exactly what you’re missing, but I recommend one of the standard analytical functionalist papers, like Lewis’s "Psychophysical and Theoretical Identifications" (Aust J Phil, 1972), or even Ryle’s _The Concept of Mind_. As for the TTT, I suggest carefully distinguishing the *conceptual* from the *empirical* dependence of mental properties on TTT-function. I take it that you accept empirical but not conceptual dependence (as you say, it’s conceivable that the TTT might be wrong). By contrast, the analytic functionalist holds that mental properties are *conceptually* dependent on causal organization -- i.e. all there is to the notion of a system’s being in a mental state is that it has a certain causal organization, and that it’s appropriately related to the environment. The standard view, I take it, is that this is an unsatisfying analysis of phenomenal mental states such as qualia, but that it goes through quite well for most other mental states, such as beliefs.

>
>sh> (3) If qualia fade and the system stays TTT-grounded, I would say
>
>sh> aboutness was gone too (what would you say, and what would it amount to
>
>sh> to be WRONG about that, even from a God’s-Eye view?)

Well, I think that it’s empirically most unlikely that qualia *would* fade, as this would mean that phenomenal states and psychological states were radically "decoherent" from each other, in a subtle sense. (I have an eternally unfinished paper, "Absent Qualia, Fading Qualia, Dancing Qualia", on just this topic.) But it’s certainly a conceptual possibility. So given this conceptual possibility, what would I say about the aboutness? I’d say that it would still be there. What would it amount to to be wrong about that? The same sort of thing it would amount to to be wrong about a system’s being alive -- e.g., that one had misanalyzed the functional capacities of the system. Aboutness is no more of an extra, free-floating fact about a system than life is.

--Dave Chalmers.

Date: Fri, 22 May 92 13:35:56 EDT From: "Stevan Harnad"

Three basic points characterize my disagreement with David Chalmers:

(1) Computational structure is not the same as causal structure. When a digital computer simulates an airplane, they are computationally equivalent but they are not causally equivalent. Causal equivalence would mean having the same causal powers, in the same "medium" (except for causally irrelevant implementational differences). An internal combustion and electric plane would be causally equivalent in their capacity to fly in the air. A simulated airplane and a real airplane are not causally equivalent but only formally equivalent (in some respects).

(2) What makes thinking different from flying is NOT that it "supervenes" on causal structure the way, say, life might, but that it is UNOBSERVABLE (or rather, observable only to the thinker). This is what allows us to forget the differences between simulated thinking and real thinking in a way that we cannot do with simulated flying and real flying.

(3) The "aboutness" of thinking is not independent of the question of qualia, it is completely parasitic on it. A system that has no qualia has no aboutness, because there is no one home in there for the symbols to be "about" anything TO.

If a system’s symbols are uninterpretable, the absence of aboutness is fairly obvious.

If a system’s symbols are systematically interpretable, as the symbols in a book or a TT-passing computer are, the lack of aboutness is less obvious, but only because of the hermeneutic power the interpretation wields over us; but this interpretability-as-being-about-something is not grounded in the book or computer but parasitic on the grounded meanings in the head of the interpreter.

If a system can pass the TTT, then its symbols are grounded, but if it still lacked qualia, it would still lack aboutness (and we would have to turn to the TTTT, according to which some TTT implementations do have minds and some don’t).

If a TTTT-indistinguishable implementation still lacked qualia, then it would still lack aboutness, only the implementations with qualia would have minds, and dualism would be correct.

>dc> It’s *computation* that’s implementation-independent, not
>dc> solar-system-hood... other implementations of that computation aren’t
>dc> solar systems.
>dc>
>dc> The strong-AI hypothesis is that unlike these properties, *cognition*
>dc> is a property that supervenes on abstract causal organization. This may
>dc> or may not be obvious at first glance, but note that unlike digestion
>dc> and solar-system-hood, it’s not ruled out at first glance: there
>dc> doesn’t seem to be any physical property independent of causal
>dc> organization that’s conceptually constitutive of cognition.

Suppose there is a Lisp program that simulates the solar system. Here is one form of implementation-independence (the kind I mean): That same program (a recipe for syntactic symbol manipulations) can be run on a Vax or a Sparc (etc.); the computations are implementation-independent and all the implementations are both formally and causally equivalent. Here is another form of "implementation-independence": This is the real solar system, that is the Lisp program running on a Sparc. They are both "performing the same computations." These two "implementations" are just formally, not causally equivalent.

The only thing that sets apart "cognition" (thinking) "at first glance" from, say, moving, is that moving is observable and thinking is not.

Fortunately, unlike in the case of, say, "living" (about which I think some of our wrong-headed intuitions may actually have been parasitic on imagining somebody being at home in there, experiencing), in the case of thinking we at least have first-person testimony (like Searle's, when he tells us we would be wrong to conclude from what we could and could not observe, that he understood Chinese) to remind us that there's more to thinking than just implemented, syntactic symbol manipulation.

(With solar systems "computing," I don't know what "computation" is any more, so let's just talk about properties of all implementations of the same syntactic symbol manipulations.)

>dc> there's a certain [real] causal structure that every
>dc> implementation of a given computation shares. That's precisely what the
>dc> definition of implementation guarantees. When a given 2-state FSA is
>dc> implemented on my computer, for instance, there are real physical
>dc> state-types in the implementation such that being in state A causes a
>dc> transition into state B, and vice versa. When a neuron-by-neuron
>dc> simulation of the brain is implemented on my computer, there are real
>dc> physical states (registers, or memory locations, or whatever) in the
>dc> implementation corresponding to the state of each neuron, and these
>dc> states interact with each other in a causal pattern isomorphic to a
>dc> pattern of interaction among the neurons.

But, alas, among the causal structures shared by the brain and the neural simulation of it is included neither the requisite causal structure for passing the TTT (observable) nor (lockstep with the former, on my hypothesis) the requisite causal structure for thinking (which happens to be unobservable to all but the [in this case nonexistent] subject).

>dc> by "causal structure" I mean, roughly, *organizational* properties of a
>dc> system: i.e., the patterns of interactions between various states,
>dc> without taking into account what those states actually are... This has
>dc> to be kept quite separate from questions about semantics.

It would help if we could speak less abstractly, resorting even to examples. There are lots of possible patterns of interaction between states. The only relevant kind for me (in discussing what I, at least, mean by computation) is the manipulation of symbols purely on the basis of their shapes, i.e., syntactically, as in a digital computer. Of course syntactic interactions are independent of semantics (although the only ones of interest are the ones that are semantically interpretable).

>dc> Given any specification of the causal structure of the brain -- even
>dc> all the way down to atoms, or whatever -- then that causal structure
>dc> could in principle be implemented in a different medium, such as
>dc> silicon. We'd just have to set it up so that our little bits of silicon
>dc> are interacting with each other according to the same patterns as the
>dc> neurons, or the atoms or whatever, were interacting with each other...
>dc> a neurophysiological identity theorist would not agree that mental
>dc> states supervene on causal structure.

Nor would I, if this were what "causal structure" was.

I have no problem with synthetic brains, made out of all kinds of unnatural parts -- as long as they retain the relevant causal powers of the brain (which for me that is just TTT-power -- for a TTTT-theorist, further neurobiological causality would matter too, perhaps even protein synthesis). I don't care what materials a computer is made of. But if all it does is manipulate symbols on the basis of syntactic rules, no matter how systematically all those syntactic goings-on can be equated with and interpretated as what goes on in the brain, nothing "mental" is "supervening" on it (because the requisite causal structure has not been duplicated).

>dc> Perhaps you don't agree that mental states depend solely on causal
>dc> structure either, because you seem to assign an essential role to I/O
>dc> transducers, and presumably it makes a difference just what kinds of
>dc> physical things -- heat, light, or whatever -- are being transduced.
>dc> Whereas a strict functionalist like myself would hold that at least
>dc> when it comes to fixing phenomenal mental states, the specific physical
>dc> nature of what's being transduced is irrelevant. On this view, a system
>dc> that merely reproduced the causal organization of the transduction in a
>dc> different medium would have the same phenomenal properties.

All I want to do is refrain from over-interpreting interpretable systems just because the thinking that they are interpretable as doing happens to be unobservable. Fortunately, the TTT, which requires real transduction (otherwise its not the TTT) is observable. Capacity for interactions with the world is hence part of the requisite causal structure for thinking.

Let me make it even simpler. Take my position to be equivalent to the hypothesis that thinking "supervenes" on BEING an optical transducer, such as a retina. There are many different kinds of optical transducer, natural and synthetic, but they must all be able to transduce real light; without that, they simply lack the requisite causal structure to be a transducer.

(Remember, you, as a computationalist [or "symbolic functionalist"] hypothesize that thinking "supervenes" on computation/TT-power alone; I, because of the symbol grounding problem, reject that and hypothesize that thinking "supervenes" only on hybrid systems with TTT-power ["robotic functionalism"], and that necessarily includes trandsuction and excludes computation alone.)

>dc> I don't think that "Is cognition computation?" is quite the right
>dc> question to ask. The right question, rather, is "Is computation
>dc> sufficient for cognition?" An advocate of strong AI might reasonably
>dc> hold that cognition in the brain is not itself computation, but that
>dc> computation is nevertheless capable of reproducing the relevant

>dc> properties (e.g. causal structure) on which cognition depends...
>dc> I certainly don’t think that the brain is a Turing machine, but I think
>dc> that nevertheless Turing machine computations are capable of cognition.
>dc> It’s a subtle point, but too often advocates of AI are saddled with
>dc> unnecessary claims such as "the brain is a computer", or "the mind is a
>dc> program".

Computation is sufficient but not necessary for cognition? I.e., the brain may not be a computer, but a computer can still think? Sounds even more far-fetched than the stronger equivalence claim -- and just as wrong, for just about the same reasons.

>dc> The question is only whether semantic content is itself *constitutive*
>dc> of something’s being a computation. To that question, the answer seems
>dc> obviously to be no. Construct an arbitrary large Turing machine by
>dc> throwing together quadruples randomly. It’s most unlikely that there
>dc> will even *be* a nontrivial semantic interpretation for this. It’s
>dc> probably not a very interesting computation, but it’s computation.
>dc>
>dc> Most interesting computations will probably turn out to have some kind
>dc> of semantic interpretation -- otherwise why would we bother with them?
>dc> ... But the notion that lies at the foundation of the computationalist
>dc> view about cognition is not "interesting computation", it’s
>dc> "computation" straight. Making some sense of the notion of "interesting
>dc> computation" is an interesting question in its own right, but it’s
>dc> independent of Searle’s original question about what makes something a
>dc> computation.

This all seems to be terminological quibbling. Whatever you want to call the rest, only interpretable computations are at issue here.

>dc> To set out the lay of the land, I agree that there’s only one mind-body
>dc> Problem worthy of a capital P, and that’s the problem of qualia. That’s
>dc> not to say that qualia are the only kind of mental states... there are
>dc> also "psychological states", those characterized by their role in the
>dc> production of behaviour rather than by their phenomenal feel. However,
>dc> there’s no more a mind-body Problem about these than there is a
>dc> "life-body Problem"... There’s no "further fact" that needs
>dc> explaining... What’s special about qualia, and makes them seem unlike
>dc> almost everything else in the world, is that there seems to be a
>dc> further fact in need of explanation, even after one has told the full
>dc> story about the mechanisms and so on.
>dc>
>dc> *All there is* to the fact of a system believing that P is that it has
>dc> the right kind of causal economy, with mechanisms that tend to produce
>dc> P-appropriate behaviour in the right sort of ways, and that are
>dc> causally related to the subject matter of P in the right sort of way.
>dc> The possession of semantic content isn’t a further fact over and above
>dc> these mechanisms: it *conceptually supervenes* on the existence of
>dc> those mechanisms, to use the philosophical parlance.

What this seems to leave out is why there should be any connection between thinking and qualia at all! It's also not clear with what justification you call beliefs "mental" or even "psychological" states. The reason there's no further fact about "aboutness" is that without qualia there's no such thing. In a system where there is nobody home, there's no one for whom anything can be "about" anything. One could still speak of grounding (TTT-power), because that, like life, does depend exclusively on observable properties. But in an insentient automaton there's simply nothing mental to speak of, be it qualitative or intentional.

>dc> Obviously, any system that is functioning in the right way is not
>dc> just "as if" alive, it's really alive, qualia or no qualia. The
>dc> same goes for belief. Maybe this means that there's not much
>dc> difference between "as if" believing and "real" believing, but why
>dc> should that bother us? We don't worry about a difference between
>dc> "as if" tables and "real" tables, after all.
>dc>
>dc> Qualia or no qualia, beliefs are still "intrinsic" (modulo questions
>dc> about narrow and wide content), in just the same way that life is
>dc> intrinsic. It's just that they're not *phenomenal*.

Terminology again. If by "intrinsic" you mean the causal substrate of beliefs is located only in the head, I agree; if you mean it's just computational, I don't. The life analogy, invoked by so many, is simply irrelevant (except inasmuch as vitalism was always parasitic, knowingly or unknowingly, on animism, as I suggested earlier).

>dc> qualia seem to be *the wrong kind of thing* to
>dc> determine that content (except perhaps for certain kinds of perceptual
>dc> content). As I said earlier, my belief about Joan of Arc may have some
>dc> associated (though hard to pin down) qualia, but it's very difficult to
>dc> see how those qualia are *constitutive* of the semantic content of the
>dc> belief. How could the *feel* of the belief possibly make it any more
>dc> about Joan of Arc than it would have been otherwise?

Because only qualia would give the beliefs a subject, a believer!

There's both a misunderstanding and an oversimplification here. TTT-grounding is what "determines the content" of thoughts, both perceptual and abstract (and it does so in a bottom-up way, so sensory grounding is primary, and higher-order concepts are grounded in lower-order ones) [see my earlier replies to Roitblat's objections to sensory bottom-uppism]. The methodological assumption is that TTT-power is sufficient for both qualia and aboutness (but this could be wrong); what's certain is that qualia are a necessary condition for aboutness: No qualia, no aboutness (just, perhaps, groundedness).

[Although Searle -- with whom I do not necessarily agree in all matters -- does not seem to have realized it yet, it is the fact that qualia are a necessary condition for aboutness that makes his own direct testimony -- to the effect that he does not understand Chinese in the Chinese Room -- both relevant and devastating to computationalism: There's something it is "like" to understand, and Searle is in a position to testify that NO SUCH THING is "supervening" on his own implementation of the TT-passing computations when he memorizes and executes the requisite symbol manipulations. The fact that some theorists, like Mike Dyer, are prepared to believe that under these conditions there would be another "system" inside Searle, one that WAS actually understanding, is just evidence of how the unobservability of understanding and the hermeneutic grip of TT-interpretability can conspire to drive one further and further into sci-fi fantasies. I think it is for similar reasons that Pat Hayes is struggling to redefine what counts as an implementation in such a way as to exclude Searle's memorizization and execution of the program.]

>dc> Your position, I take it, is roughly that: "as if" semantic content
>dc> *plus* qualia *equals* "real" semantic content. My position is that
>dc> qualia seem to contribute almost nothing to fixing the semantic content
>dc> of most beliefs, except perhaps for certain perceptual beliefs. So
>dc> whatever it is that is constitutive of "real" semantic content, qualia
>dc> don't play much of a role. This may mean that there won't be much of a
>dc> "real"/"as if" distinction to worry about (modulo the considerations
>dc> about behavioural equivalence), but that's life.

As I said, qualia don't "fix content" (except in the bottom-up sense I mentioned), but they are certainly what makes groundedness mental.

>dc> Take Joe, sitting there with some beliefs about Joan of Arc. Then a
>dc> hypothetical system (which is at least a conceptual possibility, on
>dc> your view and mine) that's physically identical to Joe but lacks
>dc> qualia, doesn't believe anything about Joan of Arc at all. I suggest
>dc> that this seems wrong. What can qualia possibly add to Joe's belief to
>dc> make them any more about Joan than they would have been otherwise?
>dc>
>dc> Well, I'm only saying that this is a *conceptual* possibility, which
>dc> surely it is on your view and mine, not an empirical possibility... But
>dc> it's entirely coherent to *imagine* a system physically identical to me
>dc> but lacking qualia... So far this view doesn't immediately imply dualism.
>dc> At least, many people who take qualia seriously accept this conceptual
>dc> possibility, but still think that ontologically, qualia aren't anything
>dc> over and above the physical... Personally, I find this view untenable...

A system in which there is no one home has no "beliefs." The conceptual possibility that the TTT may not be strong enough to guarantee that someone is home is perhaps fruitful to worry about, because it has methodological consequences, but the possibility that the TTTT would not be strong enough is not interesting (to me -- I'm not a philosopher), because it just amounts to the possibility that dualism is true.

>dc> As for the TTT, I suggest carefully distinguishing the *conceptual*
>dc> from the *empirical* dependence of mental properties on TTT-function. I
>dc> take it that you accept empirical but not conceptual dependence (as you
>dc> say, it's conceivable that the TTT might be wrong). By contrast, the
>dc> analytic functionalist holds that... all there is to the notion of a
>dc> system's being in a mental state is that it has a certain causal
>dc> organization, and that it's appropriately related to the environment...
>dc> this is an unsatisfying analysis of phenomenal mental states such as
>dc> qualia, but... goes through quite well for most other mental states,
>dc> such as beliefs.

It’s a satisfying analysis of beliefs if you forget that beliefs are supposed to have a subject.

>dc> Well, I think that it’s empirically most unlikely that qualia *would*
>dc> fade [as synthetic parts are swapped for natural ones in the brain],
>dc> as this would mean that phenomenal states and psychological states were
>dc> radically "decoherent" from each other, in a subtle sense... But it’s
>dc> certainly a conceptual possibility. So... I’d say that [aboutness]
>dc> would still be there. What would it amount to to be wrong about that?
>dc> The same sort of thing it would amount to to be wrong about a system’s
>dc> being alive -- e.g., that one had misanalyzed the functional capacities
>dc> of the system. Aboutness is no more of an extra, free-floating fact
>dc> about a system than life is.

I agree: It’s just byproduct of whatever it is that generates qualia (I’m betting on TTT-capacity).

Stevan Harnad

Date: Fri, 22 May 92 13:56:36 EDT From: "Stevan Harnad" To: harnad@gandalf.rutgers.edu Subject: Re: What is Computation?

Date: Wed, 20 May 92 22:39:48 EST From: David Chalmers

In reply to Franklin Boyle:

fb> With respect to causality, it is not enough to say just that the fb> "appropriate state-transitional relations are satisfied" [Chalmers, fb> 1992]. Rather, *how* the state-transitional relations are realized must fb> be accounted for as well. That is, *how* the physical interactions fb> among the constituent objects of the system in question actually cause fb> physical changes necessary to go from one state to the next must be fb> accounted for.

I’m open to the idea that certain constraints need to be imposed on the state-transition relations. For a start, they have to be *causal*, and there’s room for dispute over exactly what that comes to. A minimal condition is that the conditionals underwriting the relations must sustain counterfactuals (i.e., they can’t be simple material conditionals), but it’s not impossible that more is required.

One can devise some puzzle cases, where some system appears to qualify as implementing an FSA, say, under this criterion, but where one might think that it should be ruled out. For example, it turns out that an implementation of a given FSA, together with a device that simply records all inputs so far, will implement any I/O equivalent FSA (proving this is left as an exercise to the reader; the general idea is to identify the new state-types with the appropriate disjunction of states of the old system). This kind of case can probably be excluded by imposing some kind of uniformity requirement on the causal connections, but the details of this are not entirely clear to me.

That being said, I don’t find your specific argument for constraints on state-transition relations too compelling:

fb> *How* effects are brought about is important because fb> insofar as computations are processes that involve entities we hold to fb> represent (whether or not they are intrinsically referential), we have fb> to know that these representing entities are responsible for the changes fb> we observe _according_to_how_they_represent_what_they_do_ (e.g. through fb> their forms) in order to be able to call them computations in the first fb> place. Otherwise, we end up with Putnam's or Chalmers's fb> characterizations of computation, both of which are mute on the issue of fb> physical representation, even though they talk about physical states fb> (unless I'm supposed to be reading a lot more into what they're saying fb> than I am, such as unpacking the term "state correspondence" fb> [Chalmers]-- please let me know), and, therefore, admitting too many fb> systems as computational.

I agree with this, but I think that my construal of computation is capable of doing just what you say. I take the fairly standard position that representing entities represent *in virtue of their causal roles*, i.e. in virtue of the way that they affect and are affected by other states of the system, as well as the environment. According to this view, it doesn't matter precisely how the causation in question is achieved; all that matters is *that* it is achieved. Similarly, it doesn't matter *what* the internal states are that are affected; all that matters is their role in the overall causal economy of the system. So the criterion I outlined, which is silent on (a) the intrinsic nature of the states and (b) the specific manner of causation, does fine.

Of course this definition doesn't say anything explicit about reference. This is because, as I've said, I don't think that the notion of reference is conceptually prior to that of computation. Neither, for that matter, is the notion of computation prior to that of reference. Rather, I think that both of them should be analyzed in terms of the prior notion of causation. So we shouldn't expect the definition of computation (more accurately, of implementation) to say anything explicit about reference: we should simply expect that it will be *compatible* with an analysis of reference, whenever that comes along. Hopefully, given our analyses of computation and of reference, it will turn out that computational structure determines at least some aspect of representational power. The analysis of computation I've given satisfies this, being designed to be compatible with a causal-role analysis of reference.

Your view seems to be that representing entities represent by virtue of their internal form, rather than by virtue of their causal role. If this were the case, then it's possible that this construal of computation wouldn't be up to the job of fixing representational powers. However, I don't see any good reason to accept this view. I agree that the internal form of a representation can be very important -- e.g. the distributed representations in connectionist networks have complex internal structure that's central to their representational capacities. However, it seems to me that this internal form is important precisely because it *allows* a system of representations to play the kinds of causal roles that qualify them as representations. The causal role is conceptually prior, and the internal form is subsidiary.

--Dave Chalmers.

From harnad Fri May 22 14:10:29 1992 To: harnad@gandalf.rutgers.edu Subject: Re: What is Computation

Date: Thu, 21 May 1992 14:55:35 -0400 From: mcdermott-drew@CS.YALE.EDU (Drew McDermott) Subject: Second thoughts

Some second thoughts on "What is Computation?"

(1)

>
\>sh> The second concerns whether just a computer implementing a computer
>
\>sh> program can have a mind.

\>dm> I despair of ever making progress on this question without further
\>dm> empirical progress on computational modeling of thought and behavior.
\>dm> The ratio of verbiage produced to opinions changed is depressingly
\>dm> small.

I meant "depressingly large," of course.

(2) I should make it clear that all of the definitions that I (and Chalmers) have proposed are for *digital* computers. I don't think anything like that works for analog computers. Indeed, it seems as if any physical system *can* be considered to be an analog computer, in that it could be used to make predictions about the behavior of any other system modeled by the same differential equations (or whatever). (One might want to add a requirement that the inputs of the system be controllable, so that it could be used as a computer; but I wouldn't want the analogous requirement for digital computers, and there are other problems, so let's skip it.)

(3) Given our definitions, it seems obvious to me that the brain is not a digital computer -- but we shouldn't hold this against the brain. The brain's function is to control behavior and model the world. Digital technology would be the best for these purposes, but the organic world has had to make do with analog approximations to digital technology. Perhaps there is an interesting question here about when we can detect that a nondigital system is an approximation to a digital one.

Drew McDermott

From: Stevan Harnad

There are two dimensions to distinguish: (1) continuous vs. discrete and (2) "analog" vs. symbolic. The latter is, I think, the relevant distinction for this discussion. It apposes the analog world of objects (chairs, tables, airplanes, furnaces, planets, computers, transducers, animals, people) with that SUBSET of the analog world that consists of implementations of formal symbol systems, manipulating symbol tokens purely on the basis of syntax, yet interpretable as describing or representing anything else in the analog world. This has little to do, I think, with whether the brain does or does not use digital signal processing technology.

Stevan Harnad

Date: Fri, 22 May 92 10:11:15 HST From: Herbert Roitblat

Stevan Harnad wrote:

>
>sh> Three basic points characterize my disagreement with David Chalmers:

>
>sh> (1) Computational structure is not the same as causal structure. When
>
>sh> a digital computer simulates an airplane, they are computationally
>
>sh> equivalent but they are not causally equivalent. Causal equivalence
>
>sh> would mean having the same causal powers, in the same "medium" (except
>
>sh> for causally irrelevant implementational differences). An internal
>
>sh> combustion and electric plane would be causally equivalent in their
>
>sh> capacity to fly in the air. A simulated airplane and a real airplane
>
>sh> are not causally equivalent but only formally equivalent (in some
>
>sh> respects).

I think Stevan has ahold of an important relation here that needs to be amplified. An electric airplane is not a simulation of an airplane it is an implementation of an airplane that is causally equivalent with respect to flying. That is, it is equivalent to a jet plane in a limited functional domain. It would be silly for us to argue whether the electric plane is really a plane.

He goes on to say:

>
>sh> (2) What makes thinking different from flying is NOT that it
>
>sh> "supervenes" on causal structure the way, say, life might, but that it
>
>sh> is UNOBSERVABLE (or rather, observable only to the thinker). This is
>
>sh> what allows us to forget the differences between simulated thinking
>
>sh> and real thinking in a way that we cannot do with simulated flying
>
>sh> and real flying.

Here I begin to disagree. The observability of thinking relates to our ability to have knowledge about thinking, but it does not necessarily affect the causal properties of thinking. I also disagree that thinking is observable to the thinker. Even Freud argued that we do not have access to much of what we think about. Unless one defines thinking as "the narrative I" (related to the notion of subvocal speech), which by definition is narrative and accessible, thinking occurs at many different levels. Our own awareness of our cognitive processes is at best unreliable and at worst severely misleading.

More critical in this context, however, is the switch in the nature of the items being compared. One comparison is between electric planes and jet planes, the second is between simulated thinking and real thinking. Whereas it is true that the simulation is never the thing being simulated, the relation between computer based and biologically based thinking may be better characterized as like that between electric planes and jet planes than as like that between simulated planes and jet planes. Perhaps we should consider the analogical relation between planes and birds. Powered flight began in some sense as a simulation of real bird flight (e.g., DaVinci). At what point did powered flight cease to be a simulation and begin to be an implementation? Part of what I am suggesting is an investigation of the implications of assuming that computers think. If we (perhaps temporarily) entertain the assumption that computers really think, though perhaps in some computer-like way, then what do we have to conclude about thinking and computation? Are there irremediable differences between human thinking and computer thinking? The argument against computers being capable of implementing minds can be translated without loss to the argument that there are certain irremediable differences between the two. One of these is claimed to be qualia.

>
\>sh> (3) The "aboutness" of thinking is not independent of the question of
>
\>sh> qualia, it is completely parasitic on it. A system that has no qualia
>
\>sh> has no aboutness, because there is no one home in there for the symbols
>
\>sh> to be "about" anything TO.

It seems to me that the concept if qualia is entirely irrelevant to the discussion. Qualia are relics of our dualistic past. English is deeply entrenched in the folk-psychology view of dualism that our very language practically implies its validity. Qualia are category errors. The idea of qualia depends on dualistic position that someone must be home. If we irradicate dualism, then we eliminate any need for qualia. A monist has no need for qualia only sense data. If we insist on qualia, it seems to me we prejudge the question of whether computers can implement thinking, because our dualistic legacy will not permit us to entertain the notion of someone "being home," that is, the argument becomes equivalent to asking whether the computer has a soul.

Devoid of implicit dualism the notion of qualia has no more to add to the discussion than the concept of witches has to add to health and illness. Being a committed monist, I have to argue that there is no one home INSIDE me. I do not have a homunculus, or I am not a homunculus controlling a body. I think, I am, but I do not think to myself in the way that the above quotation might suggest. If there is someone home inside, then we have the familiar problem of explaining the thinking of the one inside. Does the homunculus have qualia? Does my body only simulate the intentions of the homunculus?

Herb Roitblat

Date: Mon, 25 May 92 14:54:51 EDT From: "Stevan Harnad"

METHODOLOGICAL EPIPHENOMENALISM

Herbert Roitblat wrote:

hr> The observability of thinking relates to our ability to have knowledge hr> about thinking, but it does not necessarily affect the causal hr> properties of thinking. I also disagree that thinking is observable to hr> the thinker. Even Freud argued that we do not have access to much of hr> what we think about. Unless one defines thinking as "the narrative I" hr> (related to the notion of subvocal speech), which by definition is hr> narrative and accessible, thinking occurs at many different levels. Our hr> own awareness of our cognitive processes is at best unreliable and at hr> worst severely misleading.

(1) I didn't say the observability of thinking affects the causal properties of thinking. I said there is something it's like to think. Thinking has a subject: The thinker. It is not just an insentient process that is interpretable (in its structure and its outputs) AS IF it were thinking. Hence one dead give-away of the fact that there's no thinking going on, is that there's no thinker to think them.

(2) Of course in the head of a thinker a lot is going on that he is not aware of! Why should we be aware of everything going on in our heads, or even most of it, any more than we are aware of most of what's going on outside our heads? That kind of knowledge has to be gained by honest scientific toil.

However, let's not forget that all those "unconscious thoughts" happen to be going on in the head of a conscious thinker! Forgetting this critical fact is a subtle mistake that is made over and over again, but I think that on reflection it should become obvious that it is begging the question [regarding which systems do and do not really think] to conclude that, because systems like us, that really think, have a lot going on inside them that they are not aware of, we can therefore speak of "thinking" in a system that is not aware of anything! [Or worse, that because we have an Freudian "unconscious mind" in addition to our conscious mind, other systems could have JUST an "unconscious mind"!]

Until further notice, only systems that are capable of conscious thoughts are capable of "unconscious thoughts" (which I actually think is a misnomer in any case, but that's a long story we might be able to skip for present purposes). It does not even make sense to speak of a process as "unconscious" when it's going on inside a system that has no conscious processes: Is a thermostat unconsciously thinking "It's getting hot in here"? To me, this is all just mentalistic overinterpretation.

But never mind; perhaps there are relevant similarities between what goes on in a thermostat or a computer and in my head. Fine. Let's investigate those: Maybe the similarities will turn out to yield useful generalizations, maybe not. But let's not prejudge them by assuming in advance that they are anything more than suggestive similarities. To claim that thinking is just a form of computation is precisely this kind of prejudging. If thinking were unobservable in every respect, this claim would be a normal empirical hypothesis. But since thinking IS observable to the thinker, this leaves the door to a decisive kind of negative evidence -- precisely the kind Searle used in pointing out that he would not be understanding Chinese in the Chinese Room. (By your lights, he might still be understanding it, but "unconsciously"!)

hr> More critical in this context, however, is the switch in the nature of hr> the items being compared. One comparison is between electric planes and hr> jet planes, the second is between simulated thinking and real hr> thinking. Whereas it is true that the simulation is never the thing hr> being simulated, the relation between computer based and biologically hr> based thinking may be better characterized as like that between hr> electric planes and jet planes than as like that between simulated hr> planes and jet planes. Perhaps we should consider the analogical hr> relation between planes and birds. Powered flight began in some sense hr> as a simulation of real bird flight (e.g., DaVinci). At what point did hr> powered flight cease to be a simulation and begin to be an hr> implementation?

The natural/artificial flying analogy has been invoked by computationalists many times before, and it’s just as beside the point as unconscious thoughts. I can only repeat the structure of the refutation:

(a) Unlike flying, thinking (or understanding) is unobservable (except to the thinker/understander who is doing the thinking/understanding).

(b) Searle’s Argument shows that a Chinese TT-passing system would NOT understand (and the symbol grounding problem suggests why not).

(c) Therefore, at least the stronger TTT is required in order to allow us to continue to infer that the candidate system thinks -- and this test, like a flight test, cannot be passed my computation alone. Like flying, it requires a system that is capable of transduction (at least, and probably many other analog processes as well).

(d) THIS is where the natural/artificial flying analogy IS relevant (natural thinking: ours, artificial thinking: the TTT-passing robot’s). But computation alone is no longer eligible (because of b and c).

hr> Part of what I am suggesting is an investigation of the implications of hr> assuming that computers think. If we (perhaps temporarily) entertain hr> the assumption that computers really think, though perhaps in some hr> computer-like way, then what do we have to conclude about thinking and hr> computation? Are there irremediable differences between human thinking hr> and computer thinking? The argument against computers being capable of hr> implementing minds can be translated without loss to the argument that hr> there are certain irremediable differences between the two. One of hr> these is claimed to be qualia.

If there had been no way of showing that thinking was not really going on in a computer, then the unobservability of thinking would have left this forever hopelessly underdetermined (although, unlike the underdetermination of ordinary physics, there would have been a fact of the matter: the universe as a whole contains no unobservable fact that confirms or disconfirms a Utopian physical theory that accounts for all the observables, but it does contain a fact that could disconfirm a Utopian cognitive theory -- be it a TT-, TTT-, or even TTTT-scale account of all the observable -- and that fact would be known only to the candidate system). But fortunately, in the case of computation that fact is known to us (thanks to Searle’s periscope), and the fact is that there is no Chinese-understanding going on either in Searle (unless we are prepared to believe, with Mike Dyer, that memorizing meaningless symbols can lead to multiple personality disorder) or in the computer implementing the same program he is implementing (unless we either abandon the implementation-independence of computation or Pat Hayes succeeds in finding a nonarbitrary reason for believing that Searle is not really an implementation of the same program).

So, yes, there might or might not be some helpful similarities between thinking and computation, but thinking is definitely not just computation.

hr> It seems to me that the concept of qualia is entirely irrelevant to the hr> discussion. Qualia are relics of our dualistic past. English is deeply hr> entrenched in the folk-psychology view of dualism that our very hr> language practically implies its validity. Qualia are category errors. hr> The idea of qualia depends on dualistic position that someone must be hr> home. If we eradicate dualism, then we eliminate any need for qualia. hr> A monist has no need for qualia, only sense data. If we insist on hr> qualia, it seems to me we prejudge the question of whether computers hr> can implement thinking, because our dualistic legacy will not permit us hr> to entertain the notion of someone "being home," that is, the argument hr> becomes equivalent to asking whether the computer has a soul.

The wrong-headedness of "Cartesian Dualism" is a third theme (along with unconscious thinking and natural versus artificial flying) often invoked in support of "new thinking" about cognition. I think "dualism" is being counted out prematurely, and often with insufficient understanding of just what the mind/body problem (now supposedly a non-problem) really is (was). To me it’s as simple as the intuition we all have that there is a difference between a creature that really feels it when you pinch him and another that doesn’t (because it doesn’t feel anything, no qualia, nobody home); we don’t need Descartes for that, just the experience we all share, to the effect that we really have experiences!

The presence or absence of qualia, whether or not someone is home, etc., is as relevant or irrelevant to the question of whether a system thinks or is merely interpretable as if it thinks as the presence or absence of flying is to the question of whether or not a system can fly. I would not knowingly put my money on a system that did not have qualia as a model for the mind. However, when we get past TT/computational candidates to TTT/robotic (or even TTTT/neural) candidates, where Searle’s Persiscope is no longer available, then I adopt methodological epiphenomenalism, assuming/trusting that qualia will "supervene" on the TTT-capacity, and troubling my head about them no further, since I cannot be any the wiser. What’s at issue here, however, is still pure computationalism, where one CAN indeed be the wiser, and the answer is: Nobody home.

hr> Devoid of implicit dualism the notion of qualia has no more to add to hr> the discussion than the concept of witches has to add to health and hr> illness. Being a committed monist, I have to argue that there is no one hr> home INSIDE me. I do not have a homunculus, or I am not a homunculus hr> controlling a body. I think, I am, but I do not think to myself in the hr> way that the above quotation might suggest. If there is someone home hr> inside, then we have the familiar problem of explaining the thinking of hr> the one inside. Does the homunculus have qualia? Does my body only hr> simulate the intentions of the homunculus?

And the claim that admitting that qualia/consciousness exists would lead to an infinite homuncular regress is a fourth standard canard. Sure, people have made (and continue to make) the mistake of thinking that the inputs to an organism’s brain are inputs to a homunculus inside. The best cure for this is the TTT: The system as a whole must be conscious, there has to be somebody home in there. But abandon all mentalism as soon as you address what might be going on inside the system, and concern yourself only with its capacity to generate TTT-scale performance, trusting that qualia will piggy-back on that capacity. Methodological epiphenomenalism, and no homunculus.

Stevan Harnad

Date: Sun, 24 May 92 22:24:27 -0400 From: mclennan@cs.utk.edu

Stevan,

My apologies for not replying sooner; I had to preparing a talk and attend a workshop. You wrote:

>
\>sh> (3) Searle's Chinese Room Argument and my Symbol Grounding Problem
>
\>sh> apply only to discrete symbolic computation. Searle could not implement
>
\>sh> analog computation (not even transduction) as he can symbolic
>
\>sh> computation, so his Argument would be moot against analog computation.

I'm afraid I don't understand. I see no reason why we can't have an analog version of the Chinese Room. Here it is:

Inputs come from (scaleless) moving pointers. Outputs are by twisting knobs, moving sliders, manipulating joysticks, etc. Various analog computational aids -- slide rules, nomographs, pantagraphs, etc. -- correspond to the rule book. Information may be read from the input devices and transferred to the computational aids with calipers or similar analog devices. Searle implements the analog computation by performing a complicated, ritualized sensorimotor procedure -- the point is that the performance is as mechanical and mindless as symbol manipulation. Picture an expert pilot flying an aircraft simulator. We may suppose that this analog room implements a conscious cognitive process no more farfetched than understanding Chinese, viz. recognizing a human face and responding appropriately. For concreteness we may suppose the analog computation produces the signal "Hi Granny" when presented with an image of Searle's grandmother. (My apologies to John and his grandmother.)

As in the traditional CR, the values manipulated by Searle have no *apparent* significance, except as props and constraints in his complicated dance. That is, Searle qua analog computer sees the analog values as meaningless pointer deflections and lever positions. However, with the aid of an interpreter (such as he would also need for the Chinese symbols) he might see the same analog signal as his grandmother's face.

It appears that "seeing as" is central to both the digital and analog cases. Does Searle see his performance as a syntactic (meaningless) ritual or as a semantic (meaningful) behavior? That the locus of the distinction is Searle is demonstrated by the ease with which his experience of it -- as meaningful or meaningless -- can be altered. It might be so simple as pointing out an interpretation, which would trigger a "Gestalt shift" or phenomenological reorientation, and allow these quantities and computations to be seen as saturated with meaning. Searle's experience of meaningfulness or not depends on his phenomenological orientation to the subject matter. Of course, mixed cases are also possible, as when we engage in (discrete or continuous) behaviors that have *some* significance to us, but which we don't fully understand. (Many social/cultural practices fall in this category.)

Finally, as a computer scientist devoting much of his effort to analog computation (1987, in press-a, in press-b), I am somewhat mystified by the critical distinction you draw between digital and analog computation. What convinces you that one is REAL computation, whereas the other is something else (process? pseudo-computation?)? If I'm not doing computer science please tell me what I am doing!

I hope these comments shed some light on the nature of computation (whether analog or digital), and symbol manipulation (whether discrete or continuous).

Bruce MacLennan

REFERENCES

MacLennan, B. J. (1987). Technology-independent design of neurocomputers: The universal field computer. In M. Caudill & C. Butler (Eds.), Proceedings, IEEE First International Conference on Neural Networks (Vol. 3, pp. 39-49). New York, NY: Institute of Electrical and Electronic Engineers.

MacLennan, B. J. (in press-a). Continuous symbol systems: The logic of connectionism. In Daniel S. Levine and Manuel Aparicio IV (Eds.), Neural Networks for Knowledge Representation and Inference. Hillsdale, NJ: Lawrence Erlbaum.

MacLennan, B. J. (in press-b). Characteristics of connectionist knowledge representation. Information Sciences, to appear.

Date: Mon, 25 May 92 16:59:49 EDT From: "Stevan Harnad"

ANALOG SYSTEMS AND WHAT'S RIGHT ABOUT THE SYSTEM REPLY

Bruce McLennan wrote:

bm> I see no reason why we can't have an analog version of the Chinese bm> Room. Here it is: Inputs come from (scaleless) moving pointers. Outputs bm> are by twisting knobs, moving sliders, manipulating joysticks, etc. bm> Various analog computational aids -- slide rules, nomographs, bm> pantagraphs, etc. -- correspond to the rule book. Information may be bm> read from the input devices and transferred to the computational aids bm> with calipers or similar analog devices. Searle implements the analog bm> computation by performing a complicated, ritualized sensorimotor bm> procedure -- the point is that the performance is as mechanical and bm> mindless as symbol manipulation.

I'm not sure whether you wrote this because you reject Searle's argument for the discrete symbolic case (and here wish to show that it is equally invalid for the analog case) or because you accept it for the discrete symbolic case and here wish to show it is equally valid for the analog case. In either case, I'm glad you brought it up, because it gives me the opportunity to point out exactly how simple, decisive and unequivocal my own construal of Searle's Argument is, and how clearly it applies ONLY to the discrete symbolic case:

The critical factor is the "System Reply" (the reply to the effect that it's no wonder Searle doesn't understand, he's just part of the system, and the system undersands): The refutation of the System Reply is for Searle to memorize all the symbol manipulation rules, so that the entire system that gets the inputs and generates the outputs (passing the Chinese TT) is Searle. This is how he shows that in implementing the entire symbol system, in BEING the system, he can truthfully deny that he understands Chinese. "Le Systeme, c'est Moi" is the refutation of the System Reply (unless, like Mike Dyer, you're prepared to believing that memorizing symbols causes multiple personality, or, like Pat Hayes, you're prepared to deny that Searle is really another implementation of the same symbol system the TT-passing computer implements).

But look at what you are proposing instead: You have Searle twisting knobs, using analog devices, etc. It's clear there are things going on in the room that are NOT going on in Searle. But in that case, the System Reply would be absolutely correct! I made this point explicitly in Harnad 1989 and Harnad 1991, pointing out that even an optical transducer was immune to Searle's Argument [if anyone cared to conjecture that an optical transducer could "see," in the same way it had been claimed that a computer could "understand"], because Searle could not BE another implementation of that transducer (except if he looked with his real eyes, in which case he could not deny he was seeing), whereas taking only the OUTPUT of the transducer -- as in your example -- would be subject to the System Reply. It is for this very same reason that the conventional Robot Reply to Searle misfired, because it allowed Searle to modularize the activity between a computational core, which Searle fully implemented, and peripheral devices, which he merely operated: This is why this kind of division of labor (computation doing all the real cognitive work, which is then linked to the world, like a homunculus, via trivial transducers) is such a liability to computationalism. [I've always thought computationalists were more dualistic than roboticists!]

So you have given me the chance to state again, explicitly, that Searle's Chinese Room Argument and the Symbol Grounding Problem apply ONLY to discrete formal symbol systems, in which the symbols are manipulated purely syntactically (i.e., by operations based only on the symbols' "shapes," which are arbitrary in relation to what they can be interpreted as meaning) AND where the implementation is irrelevant, i.e., where every implementation of the symbol system, despite physical differences, has the same computational properties (including the mental ones, if cognition is really computation). There is surely some implementation-independence of analog computation too (after all, there's more than one way to implement a TTT-scale robot), but that does not leave room for a Searlean implementation -- at least not one in which Searle is the entire system. Hence transduction, analog computation and the TTT are immune to Searle's Argument (as well to as the Symbol Grounding Problem, since such systems are not just implemented symbol systems in the first place).

bm> Picture an expert pilot flying an aircraft simulator. We may suppose bm> that this analog room implements a conscious cognitive process no more bm> farfetched than understanding Chinese, viz. recognizing a human face bm> and responding appropriately... As in the traditional CR, the values bm> manipulated by Searle have no *apparent* significance, except as props bm> and constraints in his complicated dance. That is, Searle qua analog bm> computer sees the analog values as meaningless pointer deflections and bm> lever positions. However, with the aid of an interpreter (such as he bm> would also need for the Chinese symbols) he might see the same analog bm> signal as his grandmother's face.

This may be true, but unfortunately it is irrelevant to anything that's at issue here (just as it's irrelevant whether Searle could eventually decrypt the Chinese symbols in the original Chinese Room). If the rules of the game allow the system to be anything but Searle himself, all bets are off, for by that token Searle could even "be" part of the real brain without understanding anything -- the brain as a whole, the "system" would be doing the understanding -- as many critics of Searle have pointed out (but for altogether the wrong reason, erroneously thinking that this fact refutes [or is even relevant to] Searle's original argument against discrete symbolic computation!). I hope this is clearer now.

bm> It appears that "seeing as" is central to both the digital and analog bm> cases. Does Searle see his performance as a syntactic (meaningless) bm> ritual or as a semantic (meaningful) behavior? That the locus of the bm> distinction is Searle is demonstrated by the ease with which his bm> experience of it -- as meaningful or meaningless -- can be altered. It bm> might be so simple as pointing out an interpretation, which would bm> trigger a "Gestalt shift" or phenomenological reorientation, and allow bm> these quantities and computations to be seen as saturated with bm> meaning. Searle's experience of meaningfulness or not depends on his bm> phenomenological orientation to the subject matter. Of course, mixed bm> cases are also possible, as when we engage in (discrete or continuous) bm> behaviors that have *some* significance to us, but which we don't fully bm> understand. (Many social/cultural practices fall in this category.)

Alas, to me, all these "Gestalt flips" are irrelevant, and the symbolic/analog distinction is the critical one.

bm> Finally, as a computer scientist devoting much of his effort to analog bm> computation (1987, in press-a, in press-b), I am somewhat mystified by the bm> critical distinction you draw between digital and analog computation. What bm> convinces you that one is REAL computation, whereas the other is something bm> else (process? pseudo-computation?)? If I'm not doing computer science bm> please tell me what I am doing! bm> bm> I hope these comments shed some light on the nature of computation bm> (whether analog or digital), and symbol manipulation (whether discrete or bm> continuous).

Unfortunately, rather than shedding light, this seems to collapse the very distinction that a logical case can be built on, independent of any mentalistic projections. I can only repeat what I wrote in response to your earlier posting (to which you have not yet replied):

> bm> a physical device is an analog computer to the extent that we
> bm> choose and intend to interpret its behavior as informing us about
> bm> some other system (real or imaginary) obeying the same formal
> bm> rules. To take an extreme example, we could use the planets as an
> bm> analog computer... >
>
>sh> (2) If all dynamical systems that instantiate differential equations
>
>sh> are computers, then everything is a computer (though, as you correctly
>
>sh> point out, everything may still not be EVERY computer, because of (1)).
>
>sh> Dubbing all the laws of physics computational ones is duly ecumenical,
>

>sh> but I am afraid that this loses just about all the special properties
>
>sh> of computation that made it attractive (to Pylyshyn (1984), for
>
>sh> example) as a candidate for capturing what it is that is special about
>
>sh> cognition and distinguishes it from from other physical processes.
>
>
>sh> (3) Searle’s Chinese Room Argument and my Symbol Grounding Problem
>
>sh> apply only to discrete symbolic computation. Searle could not implement
>
>sh> analog computation (not even transduction) as he can symbolic
>
>sh> computation, so his Argument would be moot against analog computation.
>
>sh> A grounded TTT-passing robot (like a human being and even a brain) is
>
>sh> of course an analog system, describable by a set of differential
>
>sh> equations, but nothing of consequence hangs on this level of
>
>sh> generality (except possibly dualism).

There is still the vexed question of whether or not neural nets are symbol systems. If they are, then they are subject to the symbol grounding problem. If they are not, then they are not, but then they lack the systematic semantic interpretability that Fodor & Pylyshyn (1988) have stressed as crucial for cognition. So nets have liabilities either way as long as they, like symbols, aspire to do all of cognition (Harnad 1990); in my own theory, nets play the much more circumscribed (though no less important) role of extracting the sensory invariants in the transducer projection that allow symbols to be connected to the objects they name (Harnad 1992).

Stevan Harnad

Fodor, J. & Pylyshyn, Z. (1988) Connectionism and cognitive architecture: A critical analysis. Cognition 28: 3 - 71. [also reprinted in Pinker & Mehler 1988]

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Harnad, S. (1990) Symbols and Nets: Cooperation vs. Competition. S. Pinker & J. Mehler (Eds.) (1988) "Connections and Symbols." Connection Science 2: 257-260.

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

MacLennan, B. J. (1987). Technology-independent design of neurocomputers: The universal field computer. In M. Caudill & C. Butler (Eds.), Proceedings, IEEE First International Conference on Neural Networks (Vol. 3, pp. 39-49). New York, NY: Institute of Electrical and Electronic Engineers.

MacLennan, B. J. (in press-a). Continuous symbol systems: The logic of connectionism. In Daniel S. Levine and Manuel Aparicio IV (Eds.), Neural Networks for Knowledge Representation and Inference. Hillsdale, NJ: Lawrence Erlbaum.

MacLennan, B. J. (in press-b). Characteristics of connectionist knowledge representation. Information Sciences, to appear.

Pylyshyn, Z. (1984) Computation and Cognition. Cambridge MA: MIT/Bradford

----

From: "Stevan Harnad" Date: Mon, 25 May 92 23:47:41 EDT

Date: Mon, 25 May 92 18:57:54 -0400 From: yee@envy.cs.umass.edu (Richard Yee)

SIMULTANEOUS COMPUTATIONS? (So What is a Computer?)

Can a single physical system simultaneously implement two or more non-trivial computations? Is there any significant difference between, say, a simple calculator and a (universally programmable) computer that is running a program that describes the calculator? Is a calculator a computer? Must a computer be programmable?

In an earlier posting entitled "DON'T TALK ABOUT COMPUTERS" (Apr 20, posted Apr 22), I argued that we should avoid using the term "computer" because it is ambiguous. In his reply entitled "SO WHAT IS COMPUTATION?" (Apr 22), Stevan Harnad thought I was introducing a non-standard notion of computation. He began:

>sh> Much of Yee's comment is based an a distinction between formal and
>sh> nonformal "computation," whereas my arguments are based completely on
>sh> computation as formal symbol manipulation. We will need many examples
>sh> of what nonformal computation is, plus a clear delineation of what is
>sh> NOT nonformal computation ... (It would also seem hard to
>sh> pose these questions without talking about computers, as Yee enjoins
>sh> us!)

My previous submission probably tried to juggle too many concepts in too small a space. In particular, I wanted to draw attention to TWO distinctions, not one. The first---the main subject of this message---is the distinction between *universal* and *non-universal* computation. This is entirely a standard distinction, drawn from the Theory of Computation. The second distinction is between formal and non-formal *symbol processing* and its relationship to the two types of computation. Admittedly, most of the action surrounds this question, but I will leave it for a future submission. First-things first.

Harnad concluded:

>sh> One cannot make coherent sense of this [the distinctions being made]
>sh> until the question "What is computation?", as posed in the header
>sh> to this discussion, is answered. Please reply in ordinary language
>sh> before turning again to technical formalisms, because this first pass
>sh> at formalism has merely bypassed the substantive questions that have
>sh> been raised.

OK, for the time being let us not cloud the issues with questions about formal vs. non-formal symbol processing. Unfortunately, technical formalisms remain at the heart of the matter. I think that discussions about computation should refer to the Theory of Computation (ToC). I want to argue the importance of maintaining a clear distinction between the class of Turing machines (TM's) and its PROPER SUBCLASS of universal TM's (UTM's). First, however, I will address Harnad's question about computation.

I. What is Computation?

ToC defines hierarchies of computational classes, where the most notable class is defined by the capabilities of Turing machines. I endorse the standard Church-Turing definition of computation (e.g., Lewis & Papadimitriou, 1981), which roughly states:

"computation" = what any TM does. (1)

Note carefully, it does NOT simply say:

"computation" = what any universal TM (UTM) does. (2)

Definition (1) clearly includes all UTM's, but (2) would not include all TM's. Computation must encompass what ALL TM's do, not just what universal ones do. More on this in sections II & III.

As for identifying actual instances of (approximations of?) computation in the world, I endorse most of David Chalmers' and related discussions of this subject (NB: with regard to "computation," not necessarily with regard to "beliefs," "qualia", etc.). In other words, like any theory or mathematical construct, the Turing machine model (like a triangle) is a formal abstraction of a portion of human experience. Subsequent reversing of the model (using it to view the world) involves finding physical systems that when suitably abstracted, fit the formal model (e.g., viewing New York, LA, and Miami as forming a triangle). The more constraints a model places on the world (i.e., the more *predictive* it is), the more difficult it will be to find a physical system that accidentally fits the model (i.e., that accidentally satisfies all the predictions. Try finding cities that form an isosceles right triangle, or an octagon). I take it that this, or something like it, is the "cryptographic constraint" and/or the property of "sustaining counterfactuals" which preclude arbitrary physical systems from implementing arbitrarily complex computations.

Nevertheless, a physical system will often support multiple abstract descriptions as different TM computations---just as an arrangement of cities typically can be construed as having several different (irregular) geometrical shapes. Because the human brain is a complex physical system, it undoubtedly implements numerous different TM models, all by accident. The question, of course, is whether there could exist a TM model that could both (a) "be cognitive" and (b) be implemented by the brain in a physically plausible way. Condition (a) means being a complete, verifiable, explanatory model of an individual’s cognitive processes, and condition (b) means being a causal description at the level of biological structures and processes (neurons, neurotransmitters, evolution?, etc.) and not at the levels of, say, quantum mechanics or abstract concepts (which would involve "ungrounded symbols"). The complexity entailed in condition (a) and the implementational/descriptive constraints of (b) make it vanishingly unlikely that both could be satisfied through chance.

II. The TM/UTM Distinction

Much of the symbol-grounding discussion, of course, revolves around the possibility of satisfying condition (a): Could there be a TM model that would really be cognitive if implemented (by whatever means)? Rather than trying to answer this question here, I merely want to point out one *incorrect* way to set about the task. It is incorrect, in general, to attribute properties of UNIVERSAL TM’s (computers?) to ALL TM’s.

Harnad asks:
>sh> Please give examples of what are and are
>sh> not "non-universal TM computations" and a principled explanation
>sh> of why they are or are not.

"Universal" basically means "programmable," and most TM’s are not programmable. An example of a non-universal TM is a simple calculator that performs, say, only addition. It is non-universal because one cannot give it an input (i.e., a program) that would result in its producing, say, chess-playing behaviors. [1]

What is NOT a TM computation (universal or otherwise) is a bit thornier question. TM computations include only discrete-time processes requiring a finite amount of information processing per time step. Thus, any process, e.g., an analog one, that REQUIRES (not simply "uses") infinitesimal time-steps and/or infinite-precision computations (e.g., involving irrational numbers such as pi) might be a candidate for a non-TM "computation." [2]

Computation thus includes all programmable and non-programmable TM processes, and it excludes all requirements for infinite information processing (in time or space), which would presumably exclude analog processes (if they exist).

III. Some Implications of the TM/UTM Distinction

Suppose one answered the question "What is computation?" with:

"computation" = what any computer does. (3)

This answer would be incomplete without a specific computer model. The two obvious choices are:

"computer" = TM, (4) => (1) or

"computer" = (PC, workstation, mainframe, etc.) = UTM. (5) => (2)

If you believe that *programmability* is an essential feature of computers, then you are leaning toward definition (5), and I think that many people have such a picture in mind when they talk about "computers." However, I also suspect that if forced to choose EXPLICITLY between (4) and (5), many would choose (4) because it is more general, i.e., it corresponds to definition (1), ToC's definition of computation.

But now we have a problem. Many who think and talk mainly about programmable computers, UTM's, might implicitly believe themselves to be discussing all Turing machines. They would not be. UTM's are not universal for every aspect of computation (in fact, UTM's are quite atypical). One cannot simply transform arguments regarding UTM's directly into conclusions about all TM's. THE IMPLICATION DOES NOT FLOW IN THAT DIRECTION.

The distinction between TM's and UTM's should be clear-cut. If so, then the questions posed at the beginning of this message should be answerable without philosophical debate.

Q1: Can a single physical system simultaneously implement two or more non-trivial computations? A1: Yes. Every UTM, for example, simultaneously implements a unique universal computation and another arbitrary TM computation, which depends on an input program. (It would be incoherent to recognize one computation and not the other because both satisfy exactly the same conditions for being a true implementation.)

Q2: Is there a significant difference between a simple calculator and a (universally programmable) computer (a UTM) that is running a program that describes the calculator? A2: Yes, the calculator implements ONE non-trivial computation (that we know of), while the UTM simultaneously implements TWO computations (that we know of).

Q3: Is a (simple) calculator a computer? A3: It is a TM, but not a UTM.

Q4: Must a computer be programmable? A4: A UTM must be programmable. A TM may be non-programmable, partially-programmable, or universally-programmable (a UTM).

(and continuing...)

Q5: If a UTM is running an unknown program P, can one thereby guarantee that computation X is NOT occurring? A5: No. P might describe a TM that implements X. In such a case, the physical system comprised of the UTM and its inputs, which include program P, would simultaneously implement the TM and, hence, compute X.

Q5b: What if we *knew* that X is not the UTM's computation? Could we then guarantee that X is not occurring? A5b: No, there are two computations to account for (see A5, A1).

I believe that the TM/UTM distinction has some significance for the debate surrounding the computability of mind. Some of the preceding answers cannot even be coherently formulated if one is stuck with the single term "computer." Also, the fact that the Chinese Room (CR) argument is based on a UTM should give one pause. If it is illogical to generalize from UTM's to all TM's, then does the CR say anything about the capabilities of TM's in general? More specifically, does it "prove" anything about non-programmable (rulebook-less) TM's? If one is interested in all TM's, why talk only about programmable ones? Finally, even if UTM's process symbols purely formally, can one thereby conclude that all TM's must be purely formal symbol processors?

Of course, such issues will draw us from the firm ground of the Theory of Computation, but in venturing away we should try to maintain as clear a picture as possible. Instead of thinking and talking about "computers," I think we should at least consider the ramifications of using, or of failing to use, the more-precise models of Turing machines and---when specifically intended---universal Turing machines. The TM/UTM difference is the difference between being a program and being programmed. I trust that the former is the issue of interest.

Notes:

----

[1] A TM could be partially programmable without being universal. That is, I might be able to program a calculator to compute certain formulae, e.g., celsius-fahrenheit conversions, but, again, I need not be able to make it play chess.

[2] Much of Roger Penrose's book (1989) speculates about the necessity of such infinite information-processing for mental functioning. However, it is debatable whether such processes could actually "compute something" in any proper sense, and it is not even clear that such (analog) processes actually exist.

References:

@Book{Lewis-Papadimitriou:81, author = "Lewis, H. R. and C. H. Papadimitriou", title = "Elements of the Theory of Computation", publisher = "Prentice Hall", year = "1981", address = "Englewood Cliffs, NJ" }

@Book{Penrose:89, author = "Penrose, Roger", title = "The Emperor's New Mind", publisher = "Oxford University Press", year = "1989", address = "New York", }

From: Stevan Harnad

SO WHAT IS IT THAT EVERY TM DOES?

Richard Yee has done a good job explicating the TM/UTM distinction (and has even taken some steps toward formulating an analog/TM distinction, although he thinks nothing may fit the left hand side of that distinction). I look forward to his next posting, when he tells us what it is that every TM does ["computation" = what any TM does] (and what, besides the hypothetical analog systems that he suspects may not exist) is NOT a TM, and NOT doing what every TM does, but something else (and what that something else is). (What is to be avoided in all this, I take it, is making everything a TM, and hence what everything does computation -- which, it were true, would make the statements like "Cognition is just Computation" or "The brain is just a TM" just so many tautologies.) It would also be helpful to explicate the implication-independence of computation, and just what might or might not be expected to "supervene" on it.

My hypothesis is that what TM's do is syntax: formal symbol manipulation (reading squiggles, comparing them with squoggles, erasing squiggles and replacing them by squoggles, etc.), and that whatever else might supervene on every implementation of the same computation in THIS sense, mentality does not number among them. In any case, it is only formal symbol manipulation (call it something else if "computation" is not the right word for it) that is vulnerable to Searle's

Chinese Room Argument and the Symbol Grounding Problem.

Stevan Harnad

Date: Tue, 26 May 92 12:02:50 EDT From: "Stevan Harnad"

Date: Tue, 26 May 92 10:44:40 -0400 From: davism@turing.cs.nyu.edu (Martin Davis)

Richard Yee writes emphasizing the importance of the distinction between arbitrary TMs and universal TMs.

There are some technical problems in making this distinction precise (having to do with separating the encoding of I/O data to the TM, from the actual computational process).

Long ago, I wrote on this matter:

‘‘A Note on Universal Turing Machines,’’ Automata Studies, C.E. Shannon and J. McCarthy, editors, Annals of Mathematics Studies, Princeton University Press, 1956.

‘‘The Definition of Universal Turing Machine,’’ Proceedings of the American Mathematical Society, vol.8(1957), pp. 1125-1126.

Martin Davis

Date: Tue, 26 May 92 23:36:14 EDT From: "Stevan Harnad"

Date: Tue, 26 May 1992 12:37:14 -0400 (EDT) From: Franklin Boyle

I would like to preface the following response to Dave Chalmers by repeating what Stevan, I believe, said about the What is Computation? discussion in reply to those who cautioned that it was straying from answering the original question (when it began moving towards cognition, the Chinese Room, etc.); that what is in the backs of all our minds is the desire to better understand cognition and, in particular, whether computation is sufficient to produce it.

Cognitive issues are important to this discussion because cognition is the result of a rather unique physical system (the brain) that is causally influenced by its environment as well as interactions among its consitutent parts. But unlike planetary systems and airplanes, the brain has some rather remarkable properties; in particular, its intrinsic capacity for reference. Now this is not a property of planets or solar systems or any other system we know of. So if we define computation such that planets can be construed as implementing some computation and, therefore, that they are computing, as Chalmers maintains, then we had better make sure we understand the nature of representation in such systems and how representing entities are causal, that is, how they relate to the "particular kind of state-transitional structure" [Chalmers] that serves as the basis for calling a solar system an implementation of a particular computation. Why? Because that is how the analogous argument goes for the brain as implementing a particular computation, and, thus, whether or not cognition is computation. But in this latter case, we have to account for intrinsic reference.

In other words, when we define computation, we had better be explicit about its physical characteristics because when we come to the question about the brain as a system which implements a particular computation, then whatever definition we've produced has to bear the full weight of being able to account for such things as an intrinsic referential capacity. If we allow any physical system to be an implementation of some computation, we will most likely end up with little in the way of principled criteria for determining whether cognition is computation.

>From *my* perspective, the brain is not implementing a computation just as planets in orbits are not, but for a different reason; because of structure-preserving superposition (SPS), instead of nomologically determined change (as occurs in planetary systems), as the causal mechanism for how physical change associated with its information processing is primarily brought about (see my previous postings). Both are fundamentally different from pattern matching, which, I claim, underlies computation.

Issues about consciousness, qualia, etc. should be part of another discussion on mind and brain, but symbol grounding and even the Chinese Room (unless it veers off toward arguments about multiple personalities, etc.) should be part of the "What is Computation?" discussion because they involve issues of causality and representation which are fundamental to computation. They also represent attempts to get at the kinds of things the "state transitional structure" of some computer program is going to have to support, especially in the case of the brain; e.g., "understanding" that comes about presumably because of referential characteristics of the representation.

To reiterate, if computation is going to be considered capable of producing cognition, then its state transition structure, or causal "organization", as Chalmers puts it, is going to have to explain this capacity. And, frankly, I don't believe this functionalist criterion alone can handle it. You need more than causal organization.

I therefore feel that the bifurcated discussions should, at least when the question of publication arises, be merged, with editing, under the "What is Computation?" heading.

>dc> As for continuity or discreteness, that depends on the computational
>dc> formalism one uses. Certainly all of the usual formalisms use discrete
>dc> states. Of course, a continuous physical system (like the planetary
>dc> system) can implement a discrete computation: we just have to chop up
>dc> its states in the right way (e.g. divide an orbit into 4 discrete
>dc> quadrants).

I don't believe there is a computational formalism that can legitimately be described as "computational" if it isn't discrete in a specific way. This doesn't mean that a system which is computing does not involve continuous processes (indeed, it must, if it's a physical system). But such processes are there only in a supporting capacity. They are not really part of the computation per se.

What makes computation discrete is the pattern matching process, not the nature of the representing entities. In computers, for example, symbols are discrete combinations of "high" and "low" voltages, whereas in cellular enzyme catalysis, which is also a pattern matching process (whether we call it a computation is still an issue), the tertiary structure of the enzyme onto which the subtrate molecules fit is continuous. But for both, pattern matching is a discrete event because it involves structure fitting and therefore leads to a distinct change; the switching of a particular circuit voltage from high to low or a single covalent bond, respectively. Pattee (1972) describes this type of constraint associated with structure fitting as a "decision-making" constraint. That is, the change is like a decision, which is a discrete event; a choice among alternatives.

In so-called analog computers and planetary systems, as in all other physical systems, interactions between objects cause state changes. But if you consider what is doing the representing in these two systems -- the *values* of measured attributes of the interacting objects -- you see that the representation is very different from one that is embodied by the forms of objects. Since changes in the values of measured attributes are nomologically determined, the representation in such systems not only depends on numerical values, but also on numerically based constraints (i.e., physical laws and boundary conditions) between representing (as well as nonrepresenting) entities. These are not decision-making constraints. Associations between measured attributes are not causal, yet these associations specify numerical relationships which, it would seem, would be very difficult to construe as representative of the arbitrary relationships between symbols in a pattern matching system. Structure fitting is not captured by these relationships because structures are extended, which is why they get broken up piecemeal into numerically based boundary conditions in physical state descriptions. Though this may not matter so much in the case of simple planetary systems, it does matter for cognition.

This is why you can’t just say that the orbit of a planet can be divided into 4 discrete quadrants and expect that the system is, therefore, implementing a particular computation. The causal process involved in going from one quadrant to the next is nothing like a decision-making process; it is a nomologically determined change based on Newton’s second law of motion applied to a particular system -- there is no choice among alternatives determined by the representing entities present in the system. Thus you are merely *interpreting* the system as implementing a computation because the states happen to correspond. But it is *not* an implementation (the interpretation, which is a description, can, of course, be implemented on a digitial computer. In other words, we know that the reverse is true, that a digital computer can simulate planetary motion).

>dc> Similarly for solar systems. It’s conceptually consitutive of solar-
>dc> system-hood that a system have a certain geometric shape, a certain
>dc> chemical makeup, a certain size, and so on, and these physical properties
>dc> are not determined by abstract causal organization.
>dc> .....
>dc> The strong-AI hypothesis is that unlike these properties, *cognition*
>dc> is a property that supervenes on abstract causal organization. This
>dc> may or may not be obvious at first glance, but note that unlike digestion
>dc> and solar-system-hood, it’s not ruled out at first glance: there doesn’t
>dc> seem to be any physical property independent of causal organization
>dc> that’s conceptually constitutive of cognition.

As I’ve suggested in previous posts and above, there *are* physical properties other than causal organization which, in your terminology, are conceptually consitutive of cognition-- namely, *how* cause is brought about. Why the latter constraint is "conceptually consitutive" (if I understand what you mean by this expression) of a process’s being cognition is that if the brain is to have information about objects in the world -- their structures, motions, etc. -- then it has to actually receive the projections of those objects’ structures, motions, etc. Otherwise, how could we know about them? Just saying some measured attribute or extended structure embodies it is not sufficient.

In pattern matching systems like digital computers, each structure, whether deliberately encoded by us, say as propositions, or entered directly through a video peripheral as a bitmap, causes the same behavior as long as there is a matcher to trigger the same response in both cases. Thus, it makes no difference what the structures are that do the representing in such systems. Each structure itself carries no information about its referent in pattern matching systems simply because *any* structure can cause the same outcome (of course the structures do carry information *for us* as external observers!). Furthermore, the vast amount of information implicit in the complex structures that constitute the environment cannot be just in the causal organization, because that organization is, for pattern matching systems, composed of a lot of simple actions (structureless) that trigger subsequent matches between objects (symbols) whose *actual* structures are superfluous.

This is just a physical explanation of why, as Harnad puts it, there is "nobody home" in such systems. Nor can there ever be.

>dc> To clarify, by "causal structure" I mean, roughly, *organizational*
>dc> properties of a system: i.e., the patterns of interactions between
>dc> various states, without taking into account what those states actually
>dc> are. For instance an atom, at least according to the Bohr model,
>dc> might share some causal structure with the solar system, but it differs
>dc> in many properties that aren't organizational properties, such as size,
>dc> mass, and intrinsic physical structure.

What are these "patterns of interactions between various states"? Are they just *sequences* of states or the individual interactions between particular objects that are constituents of the system? What you call "interactions between various states" are, I assume, really interactions between the constituent objects of those states, for that is what leads to new states. If it's just sequences of different states that can be mapped onto each other, without any accounting for what in those states (particular objects or their measured attributes) is actually doing the representing and whether the representing entities are causing change, then you haven't really got any principled criteria for what makes something computational.

To return to what I said in the beginning of this post about cognition, and its importance to discussions of computation; you have got to account physically for the problem of representation, since that is a fundamental part of cognition, and, therefore, should be the same for computation as well -- if you intend for computation eventually to do some cognitive work for you.

-Franklin Boyle

Pattee, H. H. (1972) Physical Problems of Decision-Making Constraints. International Journal of Neuroscience, 3:99-106.

Date: Wed, 27 May 92 22:17:19 EDT From: "Stevan Harnad"

ANALOG SYSTEMS AND THE SUBSET OF THEM THAT IMPLEMENT SYMBOL SYSTEMS

For obvious reasons, I would like to understand and agree with what Franklin Boyle has written, because, on the face of it, it looks to be among the tiny minority of contributions to this Symposium that is not in substantial disagreement with my own! Nevertheless, I have some problems with it, and perhaps Frank can help with some further clarification.

fb> unlike planetary systems and airplanes, the brain has some rather fb> remarkable properties; in particular, its intrinsic capacity for fb> reference... So if we define computation such that planets can be fb> construed as implementing some computation... as Chalmers maintains, fb> then we had better make sure we understand the nature of representation fb> in such systems and how representing entities are causal... Why? fb> Because that is how the analogous argument goes for the brain as fb> implementing a particular computation, and, thus, whether or not fb> cognition is computation. But in this latter case, we have to account fb> for intrinsic reference... If we allow any physical system to be an fb> implementation of some computation, we will most likely end up with fb> little in the way of principled criteria for determining whether fb> cognition is computation.

I agree that excessive generality about "computation" would make the question of whether cognition is computation empty, but I don’t see what THIRD possibility Frank has implicitly in mind here: For me, planets, planes, and brains are just stand-ins for ordinary analog systems. In contrast, a subset of these analog systems -- namely, computers doing computation -- are what they are, and do what they do, purely because they are implementations of the right symbol system (because they are constrained by a certain formal syntax, manipulating discrete symbols on the basis of their arbitrary shapes: "pattern matching," as Frank points out). So we have the physical analog world of objects, and some of these objects are also implementations of syntactic systems for which all specifics of the physical implementation are irrelevant, because every implementation of the same syntax is equivalent in some respect (and the respect under scrutiny here is thinking).

So I repeat, there seem to be TWO kinds of things distinguished here (actually, one kind, plus a special subset of it), namely, all physical systems, and then the subset of them that implement the same syntax, and are equivalent in that respect, independent of the physical properties of and differences among all their possible implementations. But the passage above seems to imply that there is a THIRD kind of stuff, that the brain will turn out to be that, and that that’s the right stuff (which Frank calls "intrinsic capacity for reference").

I think we differ on this, because I would distinguish only analog and syntactic systems, assuming that the relevant cognitive capacities of the brain (a hybrid nonsymbolic/symbolic system) will turn out to draw essentially on both kinds of properties, not just the syntactic properties, as the computationalists claim (and definitely not just in the sense that syntax must always have a physical implementation); and that "intrinsic reference" will turn out to be synonymous with symbol groundedness: The meanings of the internal symbols of the TTT robot will be grounded in the robot’s TTT-capacities vis-a-vis the objects, events and states of affairs to which the symbols can be systematically interpreted as referring.

fb> the brain is not implementing a computation just as planets in orbits fb> are not, but for a different reason; because of structure-preserving fb> superposition (SPS), instead of nomologically determined change (as fb> occurs in planetary systems), as the causal mechanism for how physical fb> change associated with its information processing is primarily brought fb> about (see my previous postings). Both are fundamentally different from fb> pattern matching, which, I claim, underlies computation.

I guess "SPS" is this third kind of property, but I don’t really understand how it differs from an ordinary analog process. Here’s what Frank wrote about it in his prior posting:

> fb> What other ways might physical objects cause change besides
> fb> through their [arbitrary, syntactic] forms? There are, I claim,
> fb> only two other ways: nomologically-determined change and
> fb> structure-preserving superposition (SPS). The former refers to
> fb> the kinds of changes that occur in "billiard-ball collisions".
> fb> They involve changes in the values of measured attributes
> fb> (properties whose values are numerical, such as momentum) of
> fb> interacting objects according to their pre-collisional
> fb> measured-attribute values in a physically lawful way (that is,
> fb> according to physical laws). Unlike pattern matching
> fb> interactions, these changes are not the result of structure
> fb> fitting.
> fb>
> fb> SPS is what I believe brains use. Like pattern matching (PM), it
> fb> also involves extended structure, but in a fundamentally
> fb> different way. Whereas PM involves the fitting of two
> fb> structures, which by its very nature, leads only to a simple
> fb> change such as the switching of a single voltage value from
> fb> "high" to "low" (in digital computers), SPS involves that actual
> fb> *transmission* of structure, like a stone imprinting its
> fb> structure in a piece of soft clay. That is, it is not the *form*
> fb> of a pattern or structure which must *conform* to the structure
> fb> of a matcher in order to effect system functioning (as in PM).
> fb> Rather, it is the *appearance* of that structure which causes
> fb> change because it is transmitted, so that the effect is a
> fb> structural formation of the specific features of the pattern’s
> fb> extended structure (though I won’t elaborate here, the difference
> fb> between form and appearance is somewhat akin to the difference
> fb> between the shadow of an object and the object itself). Two
> fb> different structures would physically superimpose to
> fb> automatically create a third. Harnad’s [1990] symbol grounding
> fb> processes -- "analog re-presentation" and "analog reduction" -- I
> fb> take to be examples of SPS.

Leaving out the hermeneutics of "appearance" (which I think is a dangerous red herring), the above again simply seems to be distinguishing two kinds of analog processes, but this time with the distinction mediated by properties that are interpretable as "resembling" something rather than by formal syntactic properties that are interpretable as meaning something. So, enumerating, we have (1) the usual Newtonian kind of interaction, as between planets, then we have (2) a kind of structure-preserving "impact," leaving an effect that is somehow isomorphic with its cause (like an object and its photographic image?), and then finally we have (3) implementation- independent, semantically interpretable syntactic interactions. But (2) just looks like an ordinary analog transformation, as in transduction, which I don’t think is fundamentally different from (1). In particular, if we drop talk of "appearances" and "resemblances," whatever physical connection and isomorphism is involved in (2) is, unlike (3), not merely dependent on our interpretation, hence not "ungrounded" (which is why I make extensive use of this kind of analog process in my own model for categorical perception).

My own proposal is that symbols are grounded in whatever internal structures and processes are required to generate TTT capacity, and I have no reason to believe that these consist of anything more than (1) pure analog properties, as in solar systems and their analogs, plus (2) syntactic properties, but with the latter grounded in the former, unlike in a pure (implemented but ungrounded) symbol system such as a computer. In this hybrid system (Harnad 1992 -- see excerpt below) neural nets are used to detect the invariants in the analog sensory projection that allow object categories to be connected to the symbols that name them; this model invokes no third, new property, just analog and syntactic properties.

fb> Issues about consciousness, qualia, etc. should be part of another fb> discussion on mind and brain, but symbol grounding and even the Chinese fb> Room... should be part of the "What is Computation?" discussion because fb> they involve issues of causality and representation which are fb> fundamental to computation... e.g., "understanding" ...comes about fb> presumably because of referential characteristics of the fb> representation.

But by my lights you can't partition the topic in this way, excluding the question of consciousness, because consciousness already enters as a NEGATIVE datum even in the Chinese Room: Searle testifies that he does NOT understand Chinese, therefore the implementation fails to capture intrinsic reference. Searle is reporting the ABSENCE of understanding here; that is an experiential matter. So understanding piggy-backs on the capacity to have qualia. Frank seems to agree (and to contradict this partitioning) when he writes:

fb> This is just a physical explanation of why, as Harnad puts it, fb> there is "nobody home" in such systems. Nor can there ever be.

To restate my own view, at any rate: It is an empirical hypothesis (just as computationalism, now refuted, was) that a real mind (real cognition, real thinking, somebody home, having qualia) will "supervene" on a system's TTT-capacity. Logically speaking, TTT-capacity is neither necessary nor sufficient for having a mind (this is a manifestation of the enduring -- and in my view insoluble -- mind/body problem), so the TTT-grounding hypothesis could be as wrong as computationalism. But unless someone comes up with an equivalent of Searle's Chinese Room Argument [and the ensuing Symbol Grounding Problem] against it, we can never know that the TTT hypothesis was wrong (because of the other-minds problem), so we should probably stop worrying about it. Nevertheless, it continues to be true in principle that having a real mind (somebody home, etc.) is a NECESSARY condition for the truth of the TTT hypothesis. It just happens to be a condition we will never have any way of knowing is fulfilled (which is why I am a methodological epiphenomenalist). Try denying that it's necessary without simply stipulating what thinking is by fiat (which would turn the "What is Cognition?" question into a Humpty-Dumpty matter, as empty as pan-computationalism would make the "What is Cognition?" question). To put it another way: "Groundedness = Aboutness" is just a fallible hypothesis, like any other, not a definition.

fb> You need more than causal organization.

I don't even believe syntactic systems have "causal organization" or "causal structure" in the sense Dave Chalmers claims. Their implementations of course have ordinary physical, causal properties, but these are irrelevant, by stipulation. So the only causality left is FORMAL (syntactic) causality, based on (pattern matching among) arbitrary shapes -- but with all of it being systematically interpretable as meaning something. To be sure, this is a powerful and remarkable property that symbol systems do have, but it is just a formal property, even when implemented physically. It is why nothing is really caused to move in a computer simulation of, say, the solar system or a billiard game.

fb> What makes computation discrete is the pattern matching process... fb> [which is] a discrete event because it involves structure fitting and fb> therefore leads to a distinct change; the switching of a particular fb> circuit voltage from high to low or a single covalent bond, fb> respectively. Pattee (1972) describes this type of constraint fb> associated with structure fitting as a "decision-making" constraint. fb> That is, the change is like a decision, which is a discrete event; a fb> choice among alternatives.

One can agree about the discreteness (without the unnecessary "decisional" hermeneutics), but it is still not clear what Pattee's mysterious "SPS" amounts to (although I know he invokes quantum mechanics, which I have a strong intuition is just as irrelevant as when Penrose invokes it: mysteries are not solved by applying a dose of yet another [and unrelated] mystery).

fb> In so-called analog computers and planetary systems, as in all other fb> physical systems, interactions between objects cause state changes. But fb> if you consider what is doing the representing in these two systems -- fb> the *values* of measured attributes of the interacting objects -- you fb> see that the representation is very different from one that is embodied fb> by the forms of objects. Since changes in the values of measured fb> attributes are nomologically determined, the representation in such fb> systems not only depends on numerical values, but also on numerically fb> based constraints (i.e., physical laws and boundary conditions) between fb> representing (as well as nonrepresenting) entities. These are not fb> decision-making constraints. Associations between measured attributes fb> are not causal, yet these associations specify numerical relationships fb> which, it would seem, would be very difficult to construe as fb> representative of the arbitrary relationships between symbols in a fb> pattern matching system. Structure fitting is not captured by these fb> relationships because structures are extended, which is why they get fb> broken up piecemeal into numerically based boundary conditions in fb> physical state descriptions. Though this may not matter so much in the fb> case of simple planetary systems, it does matter for cognition.

We are free to use either analog or discrete systems, natural or artificial, as tools for anything from digging a hole to reckoning time to putting a hex on an enemy. Their "representational" properties, if any, are purely extrinsic, dependent entirely on how we use them, just as the "sittability-upon" affordances of a chair are. I know that "extended structure" plays a critical role in Frank's own theory, but I have not yet been able to understand clearly what that role is. Whenever I have read about it, if I subtracted the hermeneutics, I found no remarkable property left over -- other than continuity in time and space, which is rather too general to be of any help, plus ordinary analog and syntactic interactions.

fb> Furthermore, the vast amount of information implicit in the complex fb> structures that constitute the environment cannot be just in the causal fb> organization, because that organization is, for pattern matching fb> systems, composed of a lot of simple actions (structureless) that fb> trigger subsequent matches between objects (symbols) whose *actual* fb> structures are superfluous.

Since everything seems to be susceptible to a finer-grained analysis where higher-level properties disappear, this too seems too general to be of any help in sorting out what is and is not computation or cognition.

Stevan Harnad

The following excerpt if from:

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag:

Analog Constraints on Symbols

Recall that the shapes of the symbols in a pure symbol system are arbitrary in relation to what they stand for. The syntactic rules, operating on these arbitrary shapes, are the only constraint on the manipulation of the symbols. In the kind of hybrid system under consideration here, however, there is an additional source of constraint on the symbols and their allowable combinations, and that is the nonarbitrary shape of the categorical representations that are "connected" to the elementary symbols: the sensory invariants that can pick out the object to which the symbol refers on the basis of its sensory projection. The constraint is bidirectional. The analog space of resemblances between objects is warped in the service of categorization -- similarities are enhanced and diminished in order to produce compact, reliable, separable categories. Objects are no longer free to look quite the same after they have been successfully sorted and labeled in a particular way. But symbols are not free to be combined purely on the basis of syntactic rules either. A symbol string must square not only with its syntax, but also with its meaning, i.e., what it, or the elements of which it is composed, are referring to. And what they are referring to is fixed by what they are grounded in, i.e., by the nonarbitrary shapes of the iconic projections of objects, and especially the invariants picked out by the neural net that has accomplished the categorization. If a grounding scheme like this were successful, it would be incorrect to say that the grounding was the neural net. The grounding includes, inseparably (on pain of reverting to the ungrounded symbolic circle) and nonmodularly, the analog structures and processes that the net "connects" to the symbols and vice-versa, as well as the net itself. And the system that a candidate would have to BE in order to have a mind (if this hybrid model captures what it takes to have a mind) would have to include all of the three components. Neither connectionism nor computationalism, according to this proposal, could claim hegemony in modeling cognition, and both would have to share the stage with the crucial contribution of the analog component in connecting mental symbols to the real world of objects to which they refer.

Date: Wed, 27 May 92 23:02:54 EDT From: "Stevan Harnad"

Date: Wed, 27 May 92 10:39:10 PDT From: Dr Michael G Dyer Subject: analog computation

Stevan,

Please elaborate on your reply to Bruce McLennan because I am now quite confused on just what your position is wrt intentionality and analog computation. For a moment, let's please ignore your symbol grounding issue -- i.e. let's admit that Searle, in pushing-pulling levers, etc. is doing PLENTY of "transduction" and so there IS (as you seem to require) "physical symbol grounding" (i.e. versus grounding a system within a simulated world, e.g. virtual reality systems).

Be that as it may, I thought the point of McLennan’s thought experiment was that, although Searle is moving around lots of levers and examining lots of analog dials, etc. in order to simulate some Chinese-speaking persona (i.e. the Chinese-speaking persona supervenes on Searle’s mental and physical capabilities to make the appropriate analog computations), Searle’s own subjective experience would not be anything at all like that of a person who actually understands Chinese (and can also recognize his Chinese grandmother, if you make it a TTT system).

Since Searle’s "periscope" (as you call this instrument) canNOT penetrate that Chinese mind, then WHAT makes analog computation any better (for you) than digital computation (which you claim canNOT penetrate the Chinese mind either).

To recap:

In the digital case Searle simulates the Chinese persona but Searle does not understand Chinese (nor know what the persona knows, etc.) so you and Searle conclude that there is only a simulation of understanding, not "real" understanding.

In the analog case Searle simulates the Chinese persona also and ALSO fails to understand Chinese, so again there is (from your own point of view) NO "real" understanding.

So WHAT’s so special about analog computation?

Michael Dyer

From: Stevan Harnad

Here is how I put it in Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25:

Note that the [robot] simulation/implementation distinction already points to the critical status of transduction, since Searle’s Chinese Room Argument fails completely for the robot version of the Turing Test, when the corresponding mental property at issue is the PERCEPTION of objects rather than the UNDERSTANDING of symbols. To see this, note that the terms of the Argument require Searle to show that he can take over all of the robot’s functions (thereby blocking the Systems Reply) and yet clearly fail to exhibit the mental property in question, in this case, perceiving objects. Now consider the two possible cases: (1) If Searle simulates only the symbol manipulation BETWEEN the transducers and effectors, then he is not performing all the functions of the robot (and hence it is not surprising that he does not perceive the objects the robot is supposed to perceive). (2) If, on the other hand, Searle plays homunculus for the robot, himself looking at its scene or screen, then he is BEING its transducers (and hence, not surprisingly, actually perceiving what the robot is supposed to perceive). A similar argument applies to motor activity. Robotic function, unlike symbolic function, is immune to Searle’s Chinese Room Argument.

[From Summary and Conclusions:] (7) The Transducer/Effector Argument: Prior "robot" replies to Searle have not been principled ones. They have added on robotic requirements as an arbitrary extra constraint. A principled "transducer/effector" counterargument, however, can be based on the logical fact that transduction is necessarily nonsymbolic, drawing on analog and analog-to-digital functions that can only be simulated, but not implemented, symbolically.

(8) Robotics and Causality: Searle’s argument hence fails logically for the robot version of the Turing Test, for in simulating it he would either have to USE its transducers and effectors (in which case he would not be simulating all of its functions) or he would have to BE its transducers and effectors, in which case he would indeed be duplicating their causal powers (of seeing and doing).

And here is how I put it in Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54:

Suppose that the critical question we focused on in our TTT candidate’s performance was not whether it understood symbols, but whether it could SEE. (The questions "Is it really intelligent?" "Does it really understand?" "Can it really see?" are all just variants of the question "Does it really have a mind?")...

So in the TTT variant of Searle’s thought experiment there would again be two possibilities, just as there were in the Chinese Room: In the original TT case, the machine could either really be understanding Chinese or it could just be going through the motions, manipulating symbols AS IF it understood them. Searle’s argument worked because Searle himself could do everything the machine did -- he could BE the whole system -- and yet still be obviously failing to understand.

In the TTT case of seeing, the two possibilities would again be whether the machine really saw objects or simply ACTED EXACTLY AS IF it did. But now try to run Searle’s argument through: Searle’s burden is that he must perform all the internal activities of the machine -- he must be the system -- but without displaying the critical mental function in question (here, seeing; in the old test, understanding). Now machines that behave as if they see must have sensors -- devices that transduce patterns of light on their surfaces and turn that energy into some other form (perhaps other forms of energy, perhaps symbols). So Searle seems to have two choices: Either he gets only the OUTPUT of those sensors (say, symbols), in which case he is NOT doing everything that the candidate device is doing internally (and so no wonder he is not seeing -- here the "System Reply" would be perfectly correct); or he looks directly at the objects that project onto the device’s sensors (that is, he is BEING the device’s sensors), but then he would in fact be seeing!

What this simple counterexample points out is that symbol-manipulation is not all there is to mental function, and that the linguistic version of the Turing Test just isn’t strong enough, because linguistic communication could in principle (though not necessarily in practice) be no more than mindless symbol manipulation. The robotic upgrade of the TT -- the TTT -- is hence much more realistic, because it requires the candidate to interact with the world (including ourselves) in a way that is indistinguishable from how people do it in EVERY respect: both linguistic and nonlinguistic.

The fact that mere sensory transduction can foil Searle’s argument should alert us to the possibility that sensorimotor function may not be trivial -- not just a matter of adding some simple peripheral modules (like the light-sensors that open the doors of a bank when you approach them) to a stand-alone symbol manipulator that does the real mental work. Rather, to pass the Total Turing Test, symbolic function may have to be grounded "bottom up" in NONSYMBOLIC sensorimotor function in an integrated, non-modular fashion not yet contemplated by current computer modelers.

For example, in Harnad (1987), one possible bottom-up symbol-grounding approach is described in which the elementary symbols are the names of perceptual categories that are picked out by sensory feature detectors from direct experience with objects. Nonsymbolic structures such as analog sensory projections and their invariants (and the means for learning them) would play an essential role in grounding the symbols in such a system, and the effects of the grounding would be felt throughout the system.

Harnad, S. (1987) Category induction and representation. In: Harnad, S. (Ed.) Categorical perception: The groundwork of cognition. New York: Cambridge University Press

----

Date: Wed, 27 May 92 23:16:52 EDT From: "Stevan Harnad"

Date: Wed, 27 May 1992 16:13:07 -0400 (EDT) From: Franklin Boyle

Dave Chalmers wrote (in response to my concern about representation):

>dc> I agree with this, but I think that my construal of computation is capable
>dc> of doing just what you say. I take the fairly standard position that
>dc> representing entities represent *in virtue of their causal roles*, i.e. in
>dc> virtue of the way that they affect and are affected by other states of the
>dc> system, as well as the environment. According to this view, it
>dc> doesn't matter precisely how the causation in question is achieved;
>dc> all that matters is *that* it is achieved. Similarly, it doesn't matter
>dc> *what* the internal states are that are affected; all that matters is
>dc> their role in the overall economy of the system. So the criterion I
>dc> outlined, which is silent on (a) the intrinsic nature of the states
>dc> and (b) the specific manner of causation, does fine.
>dc>
>dc> Of course this definition doesn't say anything explicit about reference.
>dc> This is because, as I've said, I don't think that the notion of reference
>dc> is conceptually prior to that of computation. Neither, for that matter,
>dc> is the notion of computation prior to that of reference. Rather, I think
>dc> that both of them should be analyzed in terms of the prior notion of
>dc> causation. So we shouldn't expect the definition of computation (more
>dc> accurately, of implementation) to say anything explicit about reference:
>dc> we should simply expect that it will be *compatible* with an analysis
>dc> of reference, whenever that comes along. Hopefully, given our analysis
>dc> of computation and of reference, it will turn out that computational
>dc> structure determines at least some aspect of representational power.
>dc> The analysis of computation I've given satisfies this, being designed
>dc> to be compatible with a causal-role analysis of reference.

In the post you're responding to, I presented (albeit briefly) an analysis of causation into three "causal mechanisms" for how all change in the world is brought about. I said that computation was founded on one of those causal mechanisms; pattern matching. In my last post, I said that reference was important if you intend to use computation to answer questions about mind, and, thus, should be factored into how you define computation. I physically restricted computation to pattern matching systems because representation/reference in such systems, e.g., digital computers, has a peculiar form-independence that SPS-based systems, which I believe underlie the brain's information processing, do not. Something has to be able to account for this difference, and I doubt that "causal role" alone is capable. Searle has already suggested that it won't.

>dc> Your view seems to be that representing entities represent by virtue
>dc> of their internal form, rather than by virtue of their causal role. If
>dc> this were the case, then it's possible that this construal of computation
>dc> wouldn't be up to the job of fixing representational powers. However,
>dc> I don't see any good reason to accept this view. I agree that the internal
>dc> form of a representation can be very important -- e.g. the distributed
>dc> representations in connectionist networks have complex internal
>dc> structure that's central to their representational capacities. However,
>dc> it seems to me that this internal form is important precisely because it
>dc> *allows* a system of representations to play the kinds of causal roles
>dc> that qualify them as representations. The causal role is conceptually
>dc> prior, and the internal form is subsidiary.

Whether representing entities represent by virtue of their internal forms or not depends on the causal mechanism. If that mechanism is pattern matching, then their forms are superfluous, so any representational aspects they have are due to causal role alone. That's just functionalism, which is certainly compatible with systems like digital computers.

This is not true for SPS (structure-preserving superposition) as the causal mechanism. In that case, the form (or, more aptly, "appearance") of the representing entity or input signal *is* the change because its structure is *transmitted*. So in your terminology, the causal role of such representing entities is not "conceptually prior" (if I'm using this expression correctly). Their extended structures (i.e., their representing aspect) *are* the changes they cause.

Let me emphasize that I am talking here about *how* objects represent, which depends on the particular causal mechanism. For example, depending on whether the causal mechanism is pattern matching or SPS, the extended structure of a representing entity is either pattern matched as a "form" or structurally transmitted, respectively, yet it may be the same physical structure in both cases. This contrasts with "type" of representation -- e.g., propositional, diagrammatic, pictorial -- which is a matter of interpretation, not causal mechanism.

-Franklin Boyle

Date: Thu, 28 May 92 15:22:44 EDT From: "Stevan Harnad"

Date: Thu, 28 May 1992 14:52:21 -0400 (EDT) From: Franklin Boyle

Stevan,

I found your points about aspects of my theory of causal mechanisms to be well taken and I plan to respond with what will hopefully be a clarification. I think (hope) you will find that my division of causal mechanisms is exactly in line with your symbol grounding through analog processes (I just happen not to call my two nonsyntactic processes "analog" because I want to distinguish them causally; I believe such a distinction is important to the issue of reference/representation). In any case, I want to give a carefully constructed reply, since your comments target the basis of my ideas. Unfortunately, I'm going out of town tomorrow afternoon for the weekend and have a zillion things to do before leaving, so I will here give only a brief reply to one point (or paragraph), both to clarify a few things and, more importantly, to ward off a potential misattribution.

>
>sh> One can agree about the discreteness (without the unnecessary
>
>sh> "decisional" hermeneutics), but it is still not clear what Pattee's
>
>sh> mysterious "SPS" amounts to (although I know he invokes
>
>sh> quantum mechanics, which I have a strong intuition is just as
>
>sh> irrelevant as when Penrose invokes it: mysteries are not solved
>
>sh> by applying a dose of yet another [and unrelated] mystery).

First, I borrowed the use of "decisional" from Pattee to describe the kind of constraints involved in structure fitting (he is primarily interested in DNA, its expression as the tertiary structures of enzymes, and issues of language codes and living systems). I assure you that I use it only as a way of describing physical processes and am careful (I hope) not to let the interpretive notions associated with the term confuse my understanding of computation and cognition (in physics, such constraints are called "non-holonomic", which Pattee also uses. But that term, I don't believe, really helps the participants in this discussion group very much).

Second, SPS (structure-preserving superposition) is my term -- I haven't seen it used anywhere else (though "superposition" is certainly prevalent -- e.g., superposition of forces in physics or superposition of activation patterns in connectionist systems). It is meant to describe my third causal mechanism. Pattee talks only about pattern matching or structure fitting, not SPS (though he may talk about the superposition of quantum states).

Third, I agree with you completely on the issue of quantum mechanics. I'm not sure exactly what Pattee's current take on this is (perhaps Eric Dietrich knows of recent publications of his on this since, I believe, they are in the same department or program -- correct me if I'm wrong), but I do know he worries about the measurement process with respect to understanding DNA and language within the context of dynamical systems; and measurement is, of course, important in quantum mechanics. But I don't think QM is relevant to the issues of computation and cognition being discussed here. And I certainly agree with your description of Penrose's analysis.

I hope this clarifies my position a little bit and the origin of SPS. I'll post a much more complete reply to your post early next week.

-Franklin Boyle

Date: Thu, 28 May 92 21:24:40 EDT From: "Stevan Harnad"

Date: Thu, 28 May 92 15:57:52 EDT From: dietrich@bingsuns.cc.binghamton.edu (dietrich)

Several times in the discussion on whether cognition is computation, someone invariably complains that certain arguments for computationalism seem to entail that everything is a computation (so it is no surprise that thinking is, too). The complainer then goes on to point out that (1) the thesis that everything is computation is vacuous, and (2) the inference from "everything is a computation" to "thinking is a computation" is also vacuous.

But neither of these claims is vacuous. The first claim is a general, universal hypothesis made in the time-honored tradition of science everywhere. The claims that everything is made of atoms, or that all objects tend to continue their motion unless otherwise disturbed, or that all species evolved are well known scientific hypotheses. They are not vacuous at all. Furthermore, the inference from, e.g., "everything is made of atoms" to "this keyboard is made of atoms" is a case of universal instantiation and constitutes a test of the hypothesis: if it should turn out that my keyboard is not made of atoms, then the hypothesis is false, and our physics is in deep trouble.

So it is with computationalism. The claim that everything is a computation is not vacuous. And the inference from it to the claim that thinking is computing is likewise not vacuous. And the further inference to my thinking is computing is a test (implicitly, anyway), because if it should turn out that my thinking is not computing (e.g., if my thinking involves executing functions that are equivalent to the halting problem or to arbitrarily large instances of the traveling salesman problem), then the claim is false. And testing for individual counterexamples is probably the only way we have of proceeding.

One can see that the claim that everything is a computation is not vacuous by noticing what it would mean if it were true and what it would take to make it false. There is an argument, by the way, that computationalism is not only not vacuous, but true. The argument is due to Chris Fields (Jetai, v.1, #3, 1989). Here is a brief synopsis.

A system is called "nonclassical" if any measurements of its internal states perturb its dynamics -- i.e., if some version of Heisenberg's principle holds for it. Given that psychological systems are nonlinear dynamical systems, it is likely that measurement perturbations of their behavior influence their future states. They are therefore, nonclassical systems. Nonclassical systems have an upper bound on the number of states we can measure, because there is an upper bound on the resolution with which states can be measured. We can detect at most a countable number of states. And this means that the behavior of the system, at the state-change level of description, can be described *completely* by a Turing machine, i.e., a partial recursive function.

(There is more to his argument, but this will do, I think.)

Fields's argument assumes that quantum mechanics is correct in its view about measurement. One can object that psychological systems (like humans) are in fact classical, but I for one don't hold out much hope for this rather desperate move.

So, here is what we have: Computationalism is not vacuous. If Fields's argument is correct, computationalism is a basic fact about the universe. Therefore, thinking is computing.

Sincerely,

Eric Dietrich

From: Stevan Harnad

Eric, I didn’t catch the part about how you would go about disconfirming the hypothesis that (1) "everything is computation" or that (2) "thinking is computation." None of that complexity-based stuff about halting problems and the describability of "nonclassical systems" sounds like a potential empirical disconfirmation to me: Obviously what can’t be computed can’t be computed; but if everything IS computation (ex hypthesi), then obviously only what can be computed is being computed (QED). And the fact that something (e.g., a "nonclassical system," to an approximation) is computationally DESCRIBABLE (SIMULABLE) is no confirmation of the fact that that something IS just (implemented, implementation-independent) computation; so it looks even less like a means of disconfirming it. By contrast, (3) "everything continues in constant motion unless disturbed" and (4) "everything is made of atoms" sound pretty readily disconfirmable to me -- they just happen to be true (on all available evidence to date).

Fortunately, the specific hypothesis that "understanding Chinese is just (implemented, implementation-independent) computation" IS disconfirmable, indeed disconfirmed, by Searle’s thought-experiment, and its failure is explained by the symbol grounding problem.

Stevan Harnad

Date: Mon, 29 Jun 92 00:02:54 EDT From: "Stevan Harnad"

Date: Wed, 17 Jun 92 13:32:03 -0400 From: mclennan@cs.utk.edu

"WORDS LIE IN OUR WAY!"

"Whenever the ancients set down a word, they believed they had made a discovery. How different the truth of the matter was! -- They had come across a problem; and while they supposed it to have been solved, they actually had obstructed its solution. -- Now in all knowledge one stumbles over rock- solid eternalized words, and would sooner break a leg than a word in doing so." -- Nietzsche (Dawn, 47)

1. THE SYSTEM REPLY

I wrote:

bm> I see no reason why we can’t have an analog version of the Chinese bm> Room. Here it is: . . . .

and Stevan replied:

>
>sh> I’m not sure whether you wrote this because you reject Searle’s argument
>
>sh> for the discrete symbolic case (and here wish to show that it is equally
>
>sh> invalid for the analog case) or because you accept it for the discrete
>
>sh> symbolic case and here wish to show it is equally valid for the analog
>
>sh> case. . . .

I was trying to argue that the analog/digital distinction could not be essential, because an analog version of the Chinese Room could be constructed, and, ceteris paribus, all arguments for or against it would still hold. I’ll address this again below.

>
>sh> The critical factor is the "System Reply" (the reply to the effect that
>
>sh> it’s no wonder Searle doesn’t understand, he’s just part of the system,
>
>sh> and the system understands): The refutation of the System Reply is for
>
>sh> Searle to memorize all the symbol manipulation rules, so that the
>
>sh> entire system that gets the inputs and generates the outputs (passing
>
>sh> the Chinese TT) is Searle. This is how he shows that in implementing
>
>sh> the entire symbol system, in BEING the system, he can truthfully deny
>
>sh> that he understands Chinese. "Le Systeme, c’est Moi" is the refutation
>
>sh> of the System Reply (unless, like Mike Dyer, you’re prepared to
>
>sh> believing that memorizing symbols causes multiple personality. . . .

Elsewhere, Stevan said:

>
>sh> . . . . If Searle memorizes all the symbols and rules, he IS the
>
>sh> system. To suppose that a second mind is generated there purely in
>
>sh> virtue of memorizing and executing a bunch of symbols and rules is (to
>
>sh> me at least) completely absurd. . . .

Well, you've forced me to blow my cover. In fact I think a version of the System Reply (the Virtual Machines Reply) is essentially correct, but I was trying to stick to the question of computation, and avoid the much- discussed issue of the System Reply and multiple minds.

But let me state briefly my position on the System Reply: If Searle could instantiate the Chinese-understanding rules, there would in fact be two minds, one (Searle's) supervening directly on the neural substrate, the other (the Chinese Understander's) supervening on Searle's rule manipulation. There is no reason to suppose that Searle would exhibit anything like a multiple personality disorder; that's a strawman. The situation is the same as a Vax running a LISP interpreter. The hardware simultaneously instantiates two interpreters, a Vax machine-code interpreter and a LISP interpreter. (N.B. The Vax is not "part" of the LISP system; it includes it all.) If we imagine that an interpreter could be aware of what it's doing, then the Vax would be aware only of interpreting Vax instructions; it would say (like Searle), "I don't know a word of LISP! How can I be understanding it? I haven't seen a stitch of LISP code; all I see are Vax instructions!" On the other hand, the LISP program is in fact being interpreted, and, under the assumption, the LISP interpreter (but not the Vax) would be aware of doing it. This may seem absurd to you, but it seems obvious to me. Let there be no mistake though: Although I take the System Reply to be valid, I do not in fact think such a set of rules (for understanding Chinese) could exist. The reason however lies elsewhere. Mais passons, indeed!

2. ANALOG COMPUTATION

Elsewhere Stevan noted:

>
>sh> There are two dimensions to distinguish: (1) continuous vs. discrete
>
>sh> and (2) "analog" vs. symbolic.

I'm glad you made this distinction, because it exposes part of the reason for our disagreement. To me the essential distinction between analog and digital computation is precisely the distinction between the continuous and the discrete. I think the terms "continuous computation" and "discrete computation" would be more accurate, but history has given us "analog computation" and "digital computation."

To avoid misunderstanding, let me point out that there is no basis to the notion that the distinction between analog and digital computation consists in the fact that analog computing is based on an "analogy" between two physical processes, whereas digital is not. (That may have been the historical origin of the terms, but now we know better.) An "analogy" between two systems is central to both kinds of computation, because in both a formal structure underlies two systems, one the computer, the other the system of interest.

Here we find exactly the syntax and semantics you have been writing about: Computation is syntactic because it it is defined in terms of formal laws referring only to physical attributes of the state, independent of its interpretation. Although computation is syntactic, semantics is also relevant because we are (mostly) concerned with systems whose states and processes (whether continuous or discrete) can be interpreted as the states and processes of some other system of interest to us. (Of course, as several others have noted, in computer science we do in fact sometimes study random programs and other programs with no intended interpretation; the reason is that we are interested in the phenomena of computation per se.)

So I suggest we purge from this discussion the terms "analog computer" and "digital computer" since they are prone to misinterpretation. If the issue is discrete vs. continuous symbols, states or processes, let's say so, and forget the rest. Henceforth I'll follow my own advice, and you'll hear no more from me about "analog" or "digital" computers (except to discuss the words).

What then is the relevant distinction? Stevan said:

>
\>sh> There are two dimensions to distinguish: (1) continuous vs. discrete
>
\>sh> and (2) "analog" vs. symbolic. The latter is, I think, the relevant
>
\>sh> distinction for this discussion. It apposes the analog world of objects
>
\>sh> (chairs, tables, airplanes, furnaces, planets, computers, transducers,
>
\>sh> animals, people) with that SUBSET of the analog world that consists of
>
\>sh> implementations of formal symbol systems, ...

It seems to me that by "the analog world" you simply mean the real world. In the real world we can distinguish (1) things (chairs, tables, airplanes, furnaces, planets, computers, transducers, animals, people) and (2) computational things, which are also part of the real world, but are important to us by virtue of instantiating certain formal processes of interest to us. The formal processes are the syntax; the semantics refers to some other process having the same formal structure. By virtue of having a semantics they are "symbolic," regardless of whether their formal structure is continuous or discrete (a matter of degree in any case, which only becomes absolute in the mathematical ideal).

3. THE GRANNY ROOM

Now let me turn to the continuous analog of the Chinese Room, which I'll dub "the Granny Room" (since its supposed purpose is to recognize the face of Searle's grandmother). My point is that there is no essential difference between the discrete and continuous cases. I wrote:

bm> I see no reason why we can't have an analog version of the Chinese bm> Room. Here it is: Inputs come from (scaleless) moving pointers. Outputs bm> are by twisting knobs, moving sliders, manipulating joysticks, etc. bm> Various analog computational aids -- slide rules, nomographs, bm> pantagraphs, etc. -- correspond to the rule book. Information may be bm> read from the input devices and transferred to the computational aids bm> with calipers or similar analog devices. . . .

Stevan replied:

>
\>sh> But look at what you are proposing instead: You have Searle twisting
>
\>sh> knobs, using analog devices, etc. It's clear there are things going on
>

>sh> in the room that are NOT going on in Searle. But in that case, the
>
>sh> System Reply would be absolutely correct! I made this point explicitly
>
>sh> in Harnad 1989 and Harnad 1991, pointing out that even an optical
>
>sh> transducer was immune to Searle’s Argument [if anyone cared to
>
>sh> conjecture that an optical transducer could "see," in the same way it
>
>sh> had been claimed that a computer could "understand"], because Searle
>
>sh> could not BE another implementation of that transducer (except if he
>
>sh> looked with his real eyes, in which case he could not deny he was
>
>sh> seeing), whereas taking only the OUTPUT of the transducer -- as in your
>
>sh> example -- would be subject to the System Reply. It is for this very
>
>sh> same reason that the conventional Robot Reply to Searle misfired,
>
>sh> because it allowed Searle to modularize the activity between a
>
>sh> computational core, which Searle fully implemented, and peripheral
>
>sh> devices, which he merely operated . . . .

Since I’m sympathetic to the System Reply, this doesn’t bother me too much, but I don’t see that the discrete Chinese Room is any more immune to it. I proposed all this apparatus to make the example (slightly) more plausible, but there is no reason it can’t all be internalized as Searle proposed in the discrete case. After all, we can do continuous spatial reasoning entirely in our heads.

Further, even if Searle memorizes all the rules, there must still be some way to get the input to him and the output from him. If a slip of paper bearing the Chinese characters is passed into the room, then he must look at it before he can apply the memorized rules; similarly he must write down the result and pass it out again. How is this different from him looking at a continuous pattern (say on a slip of paper), and doing all the rest in his head, until he draws the result on another slip of paper? Whatever you propose to do in the discrete case, I will do in the continuous. The only difference is that *inside Searle’s head* the processing will be discrete in one case and continuous in the other, but I don’t see how you can make much hang on that difference.

It seems to me that in both the discrete and continuous cases the essential point is that:

bm> . . . . the values bm> manipulated by Searle have no *apparent* significance, except as props
bm> and constraints in his complicated [mental] dance. . . .

In other words, his (mental) manipulations are purely syntactic; he's dealing with form but not content. There remains then the important question (which symbol grounding addresses) of how symbols -- whether continuous or discrete -- get their content.

4. WHAT'S A COMPUTER?

Stevan wrote:

>
>sh> (2) If all dynamical systems that instantiate differential equations
>
>sh> are computers, then everything is a computer (though, as you correctly
>
>sh> point out, everything may still not be EVERY computer, because of (1)).

I didn't say that "all dynamical systems that instantiate differential equations are computers," and certainly wouldn't conclude "everything is a computer." What I did claim was:

bm> . . . . a physical device is an analog computer to the extent that we bm> choose and intend to interpret its behavior as informing us about some bm> other system (real or imaginary) obeying the same formal rules. . . .

And later:

bm> . . . . In addition to the things that bm> are explicitly marketed as computers, there are many things that may be bm> used as computers in an appropriate context of need and availability.

That's far from saying everything is -- or even can be -- a computer! A computer, like a screwdriver, is a tool. Just as for screwdrivers, the possibility of being a computer depends both on its being physically suited to the job, as well as on its being seen as useful for the job. A knife can be a screwdriver (if we're smart enough to see it as such), but a blob of Jello cannot, no matter how creative our *seeing as*. Some physical systems can be digital (i.e., discrete) computers, others cannot; some can be analog (i.e., continuous) computers, others cannot. And most of these things will not be computers of any sort unless we see and use them as such.

>
>sh> Dubbing all the laws of physics computational ones is duly ecumenical,
>
>sh> but I am afraid that this loses just about all the special properties
>
>sh> of computation that made it attractive (to Pylyshyn (1984), for
>
>sh> example) as a candidate for capturing what it is that is special about
>
>sh> cognition and distinguishes it from from other physical processes.

True enough. But just because these are the terms in which the question has been phrased doesn't mean that they are the terms in which it can be answered. As I said:

bm> Therefore a hypothesis such as "the mind is a computer" is not bm> amenable to scientific resolution . . . . bm> . . . . A better strategy is to formulate the hypothesis in bm> terms of the notion of instantiated formal systems, which is more bm> susceptible to precise definition.

If "instantiated discrete formal system" is what we mean (or, as I would claim: instantiated formal system, whether discrete or continuous), then why don't we say so? This notion can be formally defined; "computer" and "computation" cannot, in my opinion. (Sloman, Judd and Yee have made similar suggestions.) As you said, part of the attractiveness of the computational view is a manifest constituent structure and a systematic interpretation, but this doesn't require discrete symbols, as I'll argue below. (Pace Fodor, Pylysyn et al.)

5. INTERPRETABILITY

Stevan wrote:

>
\>sh> (1) My cryptographic criterion for computerhood was not based on the
>
\>sh> uniqueness of the standard interpretation of a symbol system or the
>
\>sh> inaccessibility of nonstandard interpretations, given the standard
>
\>sh> interpretation. It was based on the relative inaccessibility
>
\>sh> (NP-Completeness?) of ANY interpretation at all, given just the symbols
>
\>sh> themselves (which in and of themselves look just like random strings of
>
\>sh> squiggles and squoggles).

This agrees with my claim above that for something to be a computer it must normally be *seen as* a computer, in other words, that its formal properties must apply to some other system of interest to us, and hence be interpretable. But I see no reason to drag in issues like NP-completeness (which probably cannot be applied rigorously in this context anyway) to impose precision on an essentially informal concept (computation). Better to talk about the relation between instantiated formal systems and their interpretations. In any case, I think the issue of interpretability (or the relative ease thereof) is irrelevant to what I take to be the substantive scientific (empirical) issue: Can cognition be adequately modeled as an instantiated discrete formal system?

6. CONTINUOUS SYMBOL SYSTEMS

>
\>sh> There is still the vexed question of whether or not neural nets are
>
\>sh> symbol systems. If they are, then they are subject to the symbol
>
\>sh> grounding problem. If they are not, then they are not, but then they
>
\>sh> lack the systematic semantic interpretability that Fodor & Pylyshyn

>
>sh> (1988) have stressed as crucial for cognition. So nets have liabilities
>
>sh> either way as long as they, like symbols, aspire to do all of cognition
>
>sh> (Harnad 1990); in my own theory, nets play the much more circumscribed
>
>sh> (though no less important) role of extracting the sensory invariants in
>
>sh> the transducer projection that allow symbols to be connected to the
>
>sh> objects they name (Harnad 1992).

All too vexed perhaps; a common consequence of asking the wrong question. We must distinguish (at least): (1) physical systems obeying differential equations, (2) continuous formal systems, and (3) continuous symbol systems (MacLennan 1988, in press-a, in press-b). We all know what class (1) is: most of the universe, so far as physics tells us. Class (2) is a subclass of class (1): systems of interest because they instantiate a given set of differential equations, but for which the actual physical quantities governed by the equations are irrelevant (that's why they're formal). (I'm glossing over the distinction between the (Platonic) abstract formal system and the (physical) instantiated formal system, but I think that's clear enough.) Continuous formal systems are treated as syntactic processes; that is, semantics is irrelevant to them qua formal system. Class (3) are those continuous formal system for which an interpretation is posited. The actual interpretation may not be specified, but we are concerned with how the continuous states and processes are related to the domain of interpretation. As noted again and again by many people, there's not much point in creating uninterpretable formal systems, so the practical distinction between (2) and (3) is whether we are interested in syntax only or syntax + semantics. (I hope the exact parallel with discrete (dynamic / formal / symbol) systems is apparent.)

As to "the vexed question of whether neural networks are symbol systems" -- it depends what you mean by neural network. Some physical systems implement Hopfield networks, but they belong in class (1), unless our interest in them consists in their implementing the abstract process, in which case they are in class (2). However, if the implemented Hopfield net refers to some other domain, perhaps an optimization problem, then it's class (3). I expect that most of the neural networks in our brains are class (3). Since class (2) is mostly of theoretical interest, it seems unlikely to be found in nature. (Of course there may be brain processes - perhaps not involving neurons at all - that nevertheless coincidentally instantiate abstract neural nets, such as Hopfield nets; these go in class (1), as do processes for which the material embodiment is critical: transducers, for example; or perhaps they are a fourth class, since they cross the 1/3 boundary. In any case symbol grounding is as relevant to continuous symbol systems as it is to discrete.)

What we normally require of discrete symbol systems, and what allows them to reduce meaningful processes to syntax, is that the interpretation be systematic, which means that it respects the constituent structure of the states. Is there anything analogous for continuous symbol systems? Indeed there is, and to find it we only need look at systematicity more abstractly. Constituent structure merely refers to the algebraic structure of the state space (e.g., as defined by the constructor operations). (There are many sources for this, but I'll take the opportunity to shamelessly plug MacLennan 1990, Chs. 2, 4.) Systematicity then simply says that the interpretation must be a homomorphism: a mapping that respects the algebraic structure (though perhaps losing some of it). The point is that these ideas are as applicable to continuous symbol systems as to the better-known discrete symbol systems. In both cases the "symbols" (physical states) are arbitrary so long as the "syntax" (algebraic structure) is preserved. If you will grant the possibility of continuous symbol systems, then I hope you will also agree that they are of critical importance to cognitive science.

REFERENCES

Fodor, J. & Pylyshyn, Z. (1988) Connectionism and cognitive architecture: A critical analysis. Cognition 28: 3 - 71. [also reprinted in Pinker & Mehler 1988]

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Harnad, S. (1990) Symbols and Nets: Cooperation vs. Competition. S. Pinker & J. Mehler (Eds.) (1988) "Connections and Symbols." Connection Science 2: 257-260.

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

MacLennan, B. J. (1988) Logic for the new AI. In J. H. Fetzer (Ed.), Aspects of Artificial Intelligence (pp. 163-192). Dordrecht: Kluwer.

MacLennan, B. J. (1990) Functional programming: Practice and theory. Reading, MA: Addison-Wesley.

MacLennan, B. J. (in press-a) Continuous symbol systems: The logic of connectionism. In Daniel S. Levine and Manuel Aparicio IV (Eds.), Neural Networks for Knowledge Representation and Inference. Hillsdale, NJ: Lawrence Erlbaum.

MacLennan, B. J. (in press-b) Characteristics of connectionist knowledge representation. Information Sciences, to appear.

Pylyshyn, Z. (1984) Computation and Cognition. Cambridge MA: MIT/Bradford.

Date: Mon, 29 Jun 92 00:04:52 EDT From: "Stevan Harnad"

WHAT ARE "CONTINUOUS SYMBOLS"?

Bruce McLennan has introduced some interesting new concepts into this discussion, in particular, the notion of continuous vs. discrete formal systems and their implementations. Some of the issues he raises are technical ones on which I do not have the expertise to render a judgment, but I think I can safely comment on the aspects of what Bruce has written that bear on what thinking can or cannot be and on what Searle’s Argument has or has not shown. I will also try to say a few words about whether there is a symbol grounding problem for "continuous symbol systems."

To summarize what I will argue below: Bruce has adopted an informal "tool" model for computers, to the effect that any system we can use to compute anything we may want to compute is a computer (including planetary systems used to compute dates, etc.). Some of these tools will compute in virtue of being programmable, discrete-state digital computers, some will compute in virtue of obeying a set of differential equations. The only restriction this places on what is or is not a computer is (1) whatever limitation there may be on what can be computed in this sense (and perhaps on what we may want to compute) and (2) whatever happen to be the properties that tools may have or lack with respect to any computation we may want to do with them.

An implication of this seems to be that (someone else's) brain is a computer to the extent that we use it (or can use it) as a tool to compute. That does not seem to be a very useful or specific conclusion; it seems rather too parasitic on how someone ELSE might use someone's brain as a tool, rather than addressing what a brain is intrinsically.

I don't find this sense of "computing" very useful (perhaps others will), nor do I find it very informative to be told that it is in this sense that the brain is really "computing" (since so much else is too, and it's hard to imagine what is not, and why not).

Now I pass to comment mode:

bm> I was trying to argue that the analog/digital distinction could not be bm> essential, because an analog version of the Chinese Room could be bm> constructed, and, ceteris paribus, all arguments for or against it would bm> still hold.

And I in turn suggested that an analog version could NOT be constructed because Searle could not implement analog "computations" all by himself the way he can implement discrete symbolic ones (his ability to do it ALL is essential, otherwise he is rightly open to the "System Reply," to the effect that it is not surprising if he does not understand, since the system as a whole understands, and he is not the whole system). I have tried to make it quite explicit that Searle's argument is valid ONLY against the hypothesis that thinking "supervenes" on any and every implementation of the right discrete symbol manipulations.

bm> Well, you've forced me to blow my cover. In fact I think a version of the bm> System Reply (the Virtual Machines Reply) is essentially correct, but I bm> was trying to stick to the question of computation, and avoid the much- bm> discussed issue of the System Reply and multiple minds. bm> bm> But let me state briefly my position on the System Reply: If Searle could bm> instantiate the Chinese-understanding rules, there would in fact be two bm> minds, one (Searle's) supervening directly on the neural substrate, the bm> other (the Chinese Understander's) supervening on Searle's rule bm> manipulation. There is no reason to suppose that Searle would exhibit bm> anything like a multiple personality disorder; that's a strawman. The bm> situation is the same as a Vax running a LISP interpreter. The hardware bm> simultaneously instantiates two interpreters, a Vax machine-code bm> interpreter and a LISP interpreter. (N.B. The Vax is not "part" of the bm> LISP system; it includes it all.) If we imagine that an interpreter could bm> be aware of what it's doing, then the Vax would be aware only of bm> interpreting Vax instructions; it would say (like Searle), "I don't know a bm> word of LISP! How can I be understanding it? I haven't seen a stitch of bm> LISP code; all I see are Vax instructions!" On the other hand, the LISP bm> program is in fact being interpreted, and, under the assumption, the LISP bm> interpreter (but not the Vax) would be aware of doing it. This may seem bm> absurd to you, but it seems obvious to me. Let there be no mistake bm> though: Although I take the System Reply to be valid, I do not in fact bm> think such a set of rules (for understanding Chinese) could exist. The bm> reason however lies elsewhere. Mais passons, indeed!

Before we pass on, let me suggest that you don't quite have the logic of Searle's Argument straight: First, if one rejects the premise that a discrete symbol system could pass the TT then one accepts Searle's conclusion that thinking is NOT just (implemented, implementation-independent, discrete) symbol manipulation. So you must at least accept the premise arguendo if you are to have anything at all to say about the validity or invalidity of Searle's argument and/or the System Reply to it.

So let's suppose, with Searle AND his opponents, that it is possible for a discrete symbol system to pass the TT; then I don't think anything of what you say above would go through. "Virtual Systems" in a discrete symbol manipulator are a HOUSE OF CARDS: If they are ungrounded at the appropriate symbolic level (the Lisp "interpreter"), then they are just as ungrounded at the lower level (the Vax machine-code "interpreter") and vice versa, because (until further notice) symbols alone are ungrounded at all "levels" (that's what's on trial here)!

The "interpreters" the machine instantiates are of course not the interpreters I'm talking about, for those "interpreters" too are merely symbols and symbol manipulations that are INTERPRETABLE as interpreters, just as the symbols at either of those virtual levels are merely interpretable as English or LISP or machine language. As hard as it is (for hermeneutic reasons) to forget it or ignore it once you've actually interpreted them, in reality, it's all just squiggles and squoggles! The highest "virtual" level is no more grounded than the lowest one. Their relation is rather like that of the double-interpretability of ACROSTICS: But apart from the interpretation WE project onto it, it's all just (systematically interpretable) syntactic gibberish -- like the symbols in a static book (in an unknown language).

That's why I keep saying that hypostasizing virtual levels (as when "we imagine that an interpreter could be aware of what it's doing") is just hermeneutics and sci-fi: Sure, the symbol system will bear the weight of the two levels of interpretation, and yes, it's uncanny that there seem to be two systematic levels of messages in there, and systematically inter-related levels to boot, as in an acrostic. But that's just what systematically interpretable symbol-manipulation is all about. Let's not make it even more mysterious than necessary in supposing that there could be an AWARENESS corresponding to the level of the LISP interpreter, for that's precisely the kind of OVERinterpretation that is ON TRIAL in the Chinese room, at any and all levels of interpretability. To try to force this supposition through as a rebuttal to Searle is just to repeat the impugned premises in a louder tone of voice!

The trouble with hermeneutics is that it makes it seem as if the shoe is on the wrong foot here: Mere interpretability-as-if, be it ever so systematic, is JUST NOT ENOUGH, irrespective of "level," and that's just what Searle's Argument demonstrates.

Now, that having been said, if what you suggest is that the TT could only be passed by what you call "continuous symbol systems," then we will have to learn more about just what continuous symbol systems are. If they should turn out to be neurons and glia and neurotransmitters, Searle would have no quarrel with that -- and even if he did, he couldn't implement them (apart from the 1st order brain he is already implementing), so his Argument would be moot. It would be equally moot if the "continuous symbol system" that could pass the TT were the planetary system, or even a simple A/A optical/acoustic transducer that transduced light intensity and spatial pattern into, say, sound intensity and some other two-dimensional analog medium. Searle could not BE an implementation of that entire system, so no wonder he lacks whatever might in reality be "supervening" on that implementation.

But I have to add that if a "continuous symbol system" rather than a discrete one could pass the TT, there would still have to be (discrete?) symbols in it corresponding to the meanings of our words and thoughts, would there not? And unless that system could also pass the TTT, those symbols would still be subject to the symbol grounding problem. Hence the system, though not penetrable by Searle's periscope, would still be ungrounded.

bm> To me the essential distinction between analog and bm> digital computation is precisely the distinction between the continuous bm> and the discrete. bm> bm> To avoid misunderstanding, let me point out that there is no basis to the bm> notion that the distinction between analog and digital computation bm> consists in the fact that analog computing is based on an "analogy" bm> between two physical processes, whereas digital is not. (That may have bm> been the historical origin of the terms, but now we know better.) An bm> "analogy" between two systems is central to both kinds of computation, bm> because in both a formal structure underlies two systems, one the bm> computer, the other the system of interest.

I agree; and what I happen to think is that rather than "analogy," what is critical here is the PHYSICAL INVERTIBILITY OF THE TRANSFORMATION FROM "OBJECT" TO "IMAGE." In discretization, some invertibility is lost; in symbolization, syntactic conventions (mediated by human interpretation) take the place of direct physical connections (except in a computer with dedicated peripherals -- an interesting and important special case). But people have had almost as much trouble with the analog/digital distinction as with defining computer/computation, and mine too are just vague intuitions, so passons...

bm> Here we find exactly the syntax and semantics you have been writing about: bm> Computation is syntactic because it is defined in terms of formal laws bm> referring only to physical attributes of the state, independent of its bm> interpretation. Although computation is syntactic, semantics is also bm> relevant because we are (mostly) concerned with systems whose states and bm> processes (whether continuous or discrete) can be interpreted as the bm> states and processes of some other system of interest to us.

I regret that I find this far too general to be helpful. If Newton's Laws are computational laws, and systems that obey them are computers, so be it, and I no longer contest that the brain (and everything else) is just a computer, doing computation. But then this generality has really said nothing of substance about anything at all: it is merely tantamount to reminding us that the nervous system, like the solar system, is a physical system, governed by natural laws that can be described formally. Who would have denied that? But this is a far cry from the claim of mentality for virtual levels of symbol interpretability in a discrete formal symbol manipulator, which is all I (and Searle) ever intended as our target.

bm> the issue is discrete vs. continuous symbols, states or processes bm> bm> It seems to me that by "the analog world" you simply mean the real world. bm> In the real world we can distinguish (1) things (chairs, tables, bm> airplanes, furnaces, planets, computers, transducers, animals, people) and bm> (2) computational things, which are also part of the real world, but are bm> important to us by virtue of instantiating certain formal processes of bm> interest to us. The formal processes are the syntax; the semantics refers bm> to some other process having the same formal structure. By virtue of bm> having a semantics they are "symbolic," regardless of whether their formal bm> structure is continuous or discrete (a matter of degree in any case, which bm> only becomes absolute in the mathematical ideal).

By the analog world I just mean physical objects and processes, whether discrete or continuous, that have whatever properties they have intrinsically, and not merely as a matter of interpretation. Some of these objects and processes can also be used to "compute" things we want to compute. Of those that are used that way, some compute in virtue of instantiating differential equations, others in virtue of instantiating discrete formal symbol manipulations. The sense of "computer" I have in mind is the latter, not the former.

Although I have not yet given it sufficient thought, it may be that even the former kind of system (which corresponds to just about anything, I should think) -- the kind of system that is an "implementation" of differential equations -- also has a kind of "symbol grounding problem," entirely independent of the question of mind-modeling, in that two radically different systems may obey formally equivalent equations (a mechanical system, say, and an electrodynamic one) and it’s just a matter of interpretation whether the terms in the equation refer to mass or charge (I’m afraid I don’t know enough physics to pick the right equivalents here). In that sense the symbols in the formal equation are "ungrounded," but as far as I know nothing much hangs on this kind of ungroundedness: On the contrary, the analogies between the different physical systems that obey equations of exactly the same form are of interest in unifying the laws of physics.

One of the features of a formal system is that you can write it down on paper. That means discrete symbols. Even the symbols of continuous differential equations are discrete. I don’t think it is appropriate to say that the continuous processes to which the symbols refer are themselves "continuous symbols" -- although I can see the motivation for this (by analogy with the implementation of discrete equations in a digital computer, in which the symbols really do correspond to binary states of the computers flip-flops).

To put it another way: I think the DISanalogies between the discrete implementation (by a computer) of the discrete symbols in a computer program and the continuous implementation (by, say, the solar system) of the discrete symbols in a set of differential equations far outweigh the analogies and are more pertinent to the question of mind-modeling and symbol grounding (and certainly to Searle’s Argument and the "Is Cognition Computation" question). But perhaps the analogies are pertinent to the "What is Computation?" question.

bm> Now let me turn to the continuous analog of the Chinese Room, which I’ll bm> dub "the Granny Room" (since its supposed purpose is to recognize the face bm> of Searle’s grandmother). My point is that there is no essential bm> difference between the discrete and continuous cases. bm> bm>
>sh> But look at what you are proposing instead: You have Searle twisting bm>
>sh> knobs, using analog devices, etc. It’s clear there are things going on bm>
>sh> in the room that are NOT going on in Searle. But in that case, the bm>
>sh> System Reply would be absolutely correct! bm> bm> Since I’m sympathetic to the System Reply, this doesn’t bother me too bm> much, but I don’t see that the discrete Chinese Room is any more immune to bm> it. I proposed all this apparatus to make the example (slightly) more bm> plausible, but there is no reason it can’t all be internalized as Searle bm> proposed in the discrete case. After all, we can do continuous spatial bm> reasoning entirely in our heads.

Now this I cannot follow at all! What on earth has "continuous spatial reasoning" in the head got to do with it? We have a robot, among whose internal functions there is, for example, transduction of light energy into some other form of energy. How is Searle to do that in its place, all in his head? In the case of the discrete symbol crunching it was quite clear what Searle had to do, and how, but what on earth are you imagining here, when you imagine him "implementing" a robot that, say, sees, partly in virtue of transducing light: How is Searle to do this without seeing, and without using any props (as he does when he memorizes all the symbol manipulation rules and processes all incoming symbols in his head)?

bm> Further, even if Searle memorizes all the rules, there must still be some bm> way to get the input to him and the output from him. If a slip of paper bm> bearing the Chinese characters is passed into the room, then he must look bm> at it before he can apply the memorized rules; similarly he must write bm> down the result and pass it out again. How is this different from him bm> looking at a continuous pattern (say on a slip of paper), and doing all bm> the rest in his head, until he draws the result on another slip of paper? bm> Whatever you propose to do in the discrete case, I will do in the bm> continuous. The only difference is that *inside Searle's head* the bm> processing will be discrete in one case and continuous in the other, but I bm> don't see how you can make much hang on that difference.

It is certainly true (and a source of much misunderstanding) that in the TT the symbols that come in are a form of input; but this (at least by my lights) is what makes the TT so equivocal, for no one supposes that the input to a real person is pure symbols: A real person, like a robot, has to have sensory transducers (optical, acoustic, or vibrotactile) to be able to perceive the sensory forms that he then interprets as linguistic symbols -- he must, in other words, have TTT capacity, even if this is not directly tested by the TT. But in the TT the problem of PERCEIVING the linguistic input is finessed: No provisions are made for TTT capacity; the symbols are simply processed "directly."

Well Searle's version of the TT simply finesses it in exactly the same way: The question of how the symbols are "perceived" is not raised, and we ask only about whether they are "understood" (the grounding problem inhering in all this should be quite evident). So we do not consider it to be a count against Searle's implementation of the entire System in the case of the TT that he did not implement the transduction of the symbols, because that is irrelevant to the TT, which is asking only about the understanding of the symbols and not their perception.

But this bracketing or modularization of perception cannot go unchallenged, and indeed the challenge is explicit in the TTT, where transduction becomes essential to the very capacity being tested: seeing.

So the answer is, for the sake of argument we agree to consider the transduction of the symbols to be trivial and modular in the case of the TT, and hence there is no "System" objection to the effect that Searle has failed to implement the transduction -- or, better, has implemented it using his own senses. But in the case of the TTT the transduction cannot be modularized without begging the question; moreover, if Searle uses his own senses he DOES see (which, if it were admitted as evidence at all -- as it probably should not be -- would have to be seen as supporting, rather than refuting, the TTT).

To put it even more briefly: sensory transduction is irrelevant to the TT but essential to the TTT; they cannot simply be equated in the two cases. And Searle simply cannot do such things "in his head."

bm> [In both the discrete and continuous case] his (mental) manipulations bm> are purely syntactic; he's dealing with form but not content. There bm> remains then the important question (which symbol grounding addresses) bm> of how symbols -- whether continuous or discrete -- get their content.

I think I agree. As I suggested earlier, whether the structures and processes inside a robot are continuous or discrete, some of them must correspond to words and thoughts, and these must be grounded (at least according to my hypothesis) in the robot's capacity to discriminate, identify and manipulate the objects, events and states of affairs that they are interpretable as being about. Otherwise they are merely dangling (even if ever so systematically) from outside interpretations.

> bm> a physical device is an analog computer to the extent that we
> bm> choose and intend to interpret its behavior as informing us about
> bm> some other system (real or imaginary) obeying the same formal rules

Fine; but in that case most things are actual or potential analog computers, including, a fortiori, me. But this definition seems to depend far too much on how we choose to USE systems, and what OTHER systems we choose to use them to explain.

> bm> In addition to the things that are explicitly marketed as computers,
> bm> there are many things that may be used as computers in an appropriate
> bm> context of need and availability.

bm> That's far from saying everything is -- or even can be -- a computer! A bm> computer, like a screwdriver, is a tool... Some physical systems can be bm> digital (i.e., discrete) computers, others cannot; some can be analog bm> (i.e., continuous) computers, others cannot. And most of these things bm> will not be computers of any sort unless we see and use them as such.

This is certainly pertinent to the "What is Computation?" discussion, but not, I think, to its underlying cognitive motivation. Also, "a physical device is an analog computer to the extent that we choose and intend to interpret its behavior as informing us about some other system (real or imaginary) obeying the same formal rules" seems to leave the doors very wide -- as wide as our imaginations.

> bm> Therefore a hypothesis such as "the mind is a computer" is not
> bm> amenable to scientific resolution . . . .

Not if we include analog computation, perhaps.

> bm> . . . . A better strategy is to formulate the hypothesis in
> bm> terms of the notion of instantiated formal systems, which is more
> bm> susceptible to precise definition. bm> bm> If "instantiated discrete formal system" is what we mean (or, as I would bm> claim: instantiated formal system, whether discrete or continuous), then bm> why don't we say so? This notion can be formally defined; "computer" and bm> "computation" cannot, in my opinion. (Sloman, Judd and Yee have made bm> similar suggestions.) As you said, part of the attractiveness of the bm> computational view is a manifest constituent structure and a systematic bm> interpretation, but this doesn't require discrete symbols, as I'll argue bm> below. (Pace Fodor, Pylysyn et al.)

An instantiated discrete formal system (systematically interpretable) is what I, at least, mean by a computer.

bm> for something to be a computer it must normally be *seen as* a bm> computer, in other words, that its formal properties must apply to some bm> other system of interest to us, and hence be interpretable.

It seems to me it was bad enough that the meanings of the symbols in the computer were just in the mind of the interpreter, but now even whether or not something is a computer is just in the mind of the interpreter. What hope has a doubly ungrounded notion like this to capture what's actually going on in the mind of the interpreter!

bm> We must distinguish (at least): (1) physical systems obeying differential bm> equations, (2) continuous formal systems, and (3) continuous symbol bm> systems (MacLennan 1988, in press-a, in press-b). We all know what class bm> (1) is: most of the universe, so far as physics tells us. Class (2) is a bm> subclass of class (1): systems of interest because they instantiate a bm> given set of differential equations, but for which the actual physical bm> quantities governed by the equations are irrelevant (that's why they're bm> formal). (I'm glossing over the distinction between the (Platonic) bm> abstract formal system and the (physical) instantiated formal system, but bm> I think that's clear enough.) Continuous formal systems are treated as bm> syntactic processes; that is, semantics is irrelevant to them qua formal bm> system. Class (3) are those continuous formal systems for which an bm> interpretation is posited. The actual interpretation may not be bm> specified, but we are concerned with how the continuous states and bm> processes are related to the domain of interpretation. As noted again and bm> again by many people, there's not much point in creating uninterpretable bm> formal systems, so the practical distinction between (2) and (3) is bm> whether we are interested in syntax only or syntax + semantics. (I hope bm> the exact parallel with discrete (dynamic / formal / symbol) systems is bm> apparent.)

I'm afraid I can't follow this. Most physical systems in the world are describable and predictable by differential equations. In that sense they are "instantiations" of those differential equations (which can also be written out on paper). We may or may not specify the intended interpretation of the formal equations as written out on paper. And we may or may not use one physical instantiation of the same set of equations to describe and predict another physical instantiation. But apart from that, what do (1) - (3) really amount to? I really don't know what a "continuous formal system" or a "continuous symbol system" is supposed to be. Equations written out on paper certainly are not continuous (the scratches on the paper are discrete symbol tokens). They may, however, correctly describe and predict continuous physical systems. That does not make those continuous physical systems either "formal" or "symbolic." In contrast, the instantiation of a discrete formal system in a digital computer running a program is indeed a discrete (implemented) formal system, because although, being physical, the computer has continuous properties too, these are irrelevant to its implementing the discrete formal system in question. And I have so far found no substantive distinction between "formal" and "symbolic."

bm> As to "the vexed question of whether neural networks are symbol systems" bm> -- it depends what you mean by neural network. Some physical systems bm> implement Hopfield networks, but they belong in class (1), unless our bm> interest in them consists in their implementing the abstract process, in bm> which case they are in class (2). However, if the implemented Hopfield bm> net refers to some other domain, perhaps an optimization problem, then bm> it's class (3). I expect that most of the neural networks in our brains bm> are class (3). Since class (2) is mostly of theoretical interest, it bm> seems unlikely to be found in nature. (Of course there may be brain bm> processes - perhaps not involving neurons at all - that nevertheless bm> coincidentally instantiate abstract neural nets, such as Hopfield nets; bm> these go in class (1), as do processes for which the material embodiment bm> is critical: transducers, for example; or perhaps they are a fourth bm> class, since they cross the 1/3 boundary. In any case symbol grounding is bm> as relevant to continuous symbol systems as it is to discrete.)

I am still unsure whether "continuous symbol systems" do or do not have a symbol grounding problem; in fact, I'm still not sure what a continuous symbol system is. And as I suggested earlier, the fact that something can be (1), (2) or (3) depending on what use we happen to choose to use them for does not seem to be a very helpful fact. Surely the brain is what it is irrespective of what we (outsiders) may want to use it for. What we are looking for (as with everything else) is the CORRECT description.

bm> What we normally require of discrete symbol systems, and what allows them bm> to reduce meaningful processes to syntax, is that the interpretation be bm> systematic, which means that it respects the constituent structure of the bm> states. Is there anything analogous for continuous symbol systems? bm> Indeed there is, and to find it we only need look at systematicity more bm> abstractly. Constituent structure merely refers to the algebraic bm> structure of the state space (e.g., as defined by the constructor bm> operations). (There are many sources for this, but I'll take the bm> opportunity to shamelessly plug MacLennan 1990, Chs. 2, 4.) Systematicity bm> then simply says that the interpretation must be a homomorphism: a bm> mapping that respects the algebraic structure (though perhaps losing some bm> of it). The point is that these ideas are as applicable to continuous bm> symbol systems as to the better-known discrete symbol systems. In both bm> cases the "symbols" (physical states) are arbitrary so long as the bm> "syntax" (algebraic structure) is preserved. If you will grant the bm> possibility of continuous symbol systems, then I hope you will also bm> agree that they are of critical importance to cognitive science.

I am agnostic about continuous symbol systems (in part because I am not competent to evaluate the technical point you make above). If there is a generalization of discrete formal symbols and symbol manipulations to continuous formal symbols and symbol manipulations with constituent structure, compositionality and systematicity (including systematic interpretability) that is useful and predictive, I could of course have no objections. The only question I would be inclined to raise concerns the "language of thought" notion that motivated proposing discrete symbols and symbol strings as a theory of mental states in the first place: The symbols in a language are, I think, necessarily discrete. What would a continuous candidate look like? As I formulated it, the symbol grounding problem is very much linked to the notion of discrete symbols in a language of thought. It is they who are ungrounded in a computer implementation. It is not even clear to me how to pose the question of groundedness for "continuous symbols."

Stevan Harnad

Date: Mon, 29 Jun 92 15:16:57 EDT From: "Stevan Harnad"

Date: Fri, 5 Jun 92 02:28:13 EST From: David Chalmers

Thanks to Franklin Boyle for his thoughtful replies. I confess to not fully understanding his position, but as far as I understand it, I gather that he's saying that for the purposes of determining what's computation and cognition, we have to look more closely than simply at the causal state-transitional structure. What matters isn't just the pattern of transitions, it's (a) the specific nature of the state-transitions, and (b) the specific nature of the causal relations.

As far as (a) is concerned, we have to distinguish between real changes in structure, e.g. "receiving a projection of a structure", from mere changes in the "measured attributes" of a system (e.g. voltages). As far as (b) is concerned, we have to distinguish between "structure-preserving superposition", in which the form (or appearance) of one state somehow imprints itself on another, from mere "pattern matching" and "structure fitting". The wrong kinds of state-transition and causation give you computation; the right kinds might give you cognition.

My reply to this proposal is pretty simple: I'm not sure that these distinctions come to anything, and I see no reason why they should make a difference between cognition and non-cognition. It seems to me that the form or appearance that various states embody is just irrelevant to a system's status as cognitive, or as computational. We could probably make an implementation of a Turing machine out of plasticine, where lumps corresponding to "symbols" change shape by colliding up against each other; it would still be a computation. And while we don't know how neural causation works, it doesn't seem entirely implausible that the basis is in the transmission of information via various "measured attributes" not unlike voltage: e.g. potentials and firing frequencies.

I've probably misunderstood this position completely, but it seems to me that however these distinctions are drawn, there's no principled reason why computation or cognition should lie on only one side of the line. (Well, maybe there's one principled reason: if the Chinese room argument were valid, there might be a motivation for a line like this. But of course the Chinese room argument isn't valid :-) .)

In reply to some more specific points:

fb> If we allow any physical system to be an implementation of fb> some computation, we will most likely end up with little in the way of fb> principled criteria for determining whether cognition is computation.

Let's dispose of this canard for once and for all. Even if every system implements some computation, this doesn't imply that every system is engaged in cognition, for the simple reason that only *certain kinds* of computation qualify as cognition. Not even the strongest of believers in strong AI has said that implementing *any* program is sufficient for cognition. It has to be the right kind of program (or, more generally, the right kind of computation).

So: just because the solar system implements a trivial 4-state FSA, we don't suddenly have an interplanetary mind: 4-state FSAs aren't the kinds of things that *think*. Isolating those kinds of computation that qualify as cognition is an interesting, highly non-trivial question in its own right. Presumably, only certain highly complex computations will be sufficient for cognition; solar systems, along with rocks and most everything else in the world, won't have the requisite causal structure to qualify.

What would make the notion of computation vacuous would be is every system implemented *every* computation. But that's just not the case.

fb> I don't believe there is a computational formalism that can fb> legitimately be described as "computational" if it isn't discrete in a fb> specific way. This doesn't mean that a system which is computing does fb> not involve continuous processes (indeed, it must, if it's a physical fb> system). But such processes are there only in a supporting capacity. fb> They are not really part of the computation per se.

I disagree with this. See the work of Bruce McLennan. With the usual variety of computation, we specify the causal patterns between discrete state-transitions via some formalism, with appropriate implementation conditions; one can do precisely the same thing for patterns of continuous state-transitions. It's just that the formalism will be something more reminiscent of differential equations than Boolean logic. However, we should probably stick with discrete computation for the purposes of this discussion.

fb> This is why you can't just say that the orbit of a planet can be fb> divided into 4 discrete quadrants and expect that the system is, fb> therefore, implementing a particular computation. The causal process fb> involved in going from one quadrant to the next is nothing like a fb> decision-making process; it is a nomologically determined change based fb> on Newton's second law of motion applied to a particular system -- fb> there is no choice among alternatives determined by the representing fb> entities present in the system.

I said earlier that the solar system is probably a bad example, as it has no counterfactual sensitivity to various inputs; and this is what you seem to be worrying about here. If we leave out sensitivity to inputs, *every* implementation of a computation undergoes nomologically determined change; it doesn't have any choice (unless we're talking about nondeterministic computation, of course).

fb> As I've suggested in previous posts and above, there *are* physical fb> properties other than causal organization which, in your terminology, fb> are conceptually consitutive of cognition-- namely, *how* cause is fb> brought about. Why the latter constraint is "conceptually consitutive" fb> (if I understand what you mean by this expression) of a process's being fb> cognition is that if the brain is to have information about objects in fb> the world -- their structures, motions, etc. -- then it has to actually fb> receive the projections of those objects' structures, motions, etc. fb> Otherwise, how could we know about them? Just saying some measured fb> attribute or extended structure embodies it is not sufficient.

There's some kind of strange, deeply-embedded assumption here: that true "knowledge" requires embedding of an object's actual "structure" inside a cognitive system. To think about spheres, does one need something spherical inside one's head? Surely not. This sounds like the kind of scholastic theory that Cummins dismisses in the first few pages of his book. Even if you don't mean something quite as literal as this, there seems to me to be nothing wrong in saying that the brain embodies the information it carries in mere "measured attributes" such as potentials, frequencies, and so on, as long as these bear the requisite causal relation to the outside world and play the appropriate functional role within the system.

fb> What are these "patterns of interactions between various states"? Are fb> they just *sequences* of states or the individual interactions between fb> particular objects that are constituents of the system? What you call fb> "interactions between various states" are, I assume, really fb> interactions between the constituent objects of those states, for that fb> is what leads to new states. If it's just sequences of different states fb> that can be mapped onto each other, without any accounting for what in fb> those states (particular objects or their measured attributes) is fb> actually doing the representing and whether the representing entities fb> are causing change, then you haven't really got any principled criteria fb> for what makes something computational.

"Sequences" is on the right track, except (a) we need a lot more than a single "sequence" -- we need to specify the different "sequences" that will arise e.g. for different inputs; (b) the states here needn't be monadic, as in simple FSAs; the overall state at a given time may be combinatorially structured (as e.g. in a Turing machine, or a neural network), with lots of substates to a given state (e.g. the state of the brain at a given time can be looked at as the combination of a lot of substates, e.g. the states of individual neurons); the causal structure of the system will then depend on the state-transitions between the substates -- it's insufficient in general to describe the system by a simple sequence of monadic states; (c) the relation between consecutive items in a "sequence" must be *causal*.

Look at the interactions as between the "constituent objects" of the states, rather than between the states themselves, if you like; it doesn't make any difference. On the view that I'm taking, it doesn't matter what a particular state corresponds to physically -- a "measured attribute" or a particular "form" or whatever -- as long as there's a fact of the matter about whether the system is in a given state, and as long as these states have the right overall pattern of causation between them.

--Dave Chalmers.

Date: Mon, 29 Jun 92 15:15:56 EDT From: "Stevan Harnad" Subject: Re: What is Computation?

[Apologies for the delay in posting this; the prior posting from Bruce McLennan was actually received earlier, then inadvertently erased, so it had to be requested again. Hence the apparent nonconsecutive order of the postings. -- SH]

Date: Tue, 2 Jun 1992 17:00:04 -0400 (EDT) From: Franklin Boyle

Stevan Harnad writes:

>
\>sh> I agree that excessive generality about "computation" would make
>
\>sh> the question of whether cognition is computation empty, but I
>
\>sh> don't see what THIRD possibility Frank has implicitly in mind
>
\>sh> here: For me, planets, planes, and brains are just stand-ins for
>
\>sh> ordinary analog systems. In constrast, a subset of these analog
>
\>sh> systems -- namely, computers doing computation -- are what they
>

>sh> are, and do what they do, purely because they are implementations
>
>sh> of the right symbol system (because they are constrained by a
>
>sh> certain formal syntax, manipulating discrete symbols on the basis
>
>sh> of their arbitrary shapes: "pattern matching," as Frank points out).
>
>sh> So we have the physical analog world of objects, and some of
>
>sh> these objects are also implementations of syntactic systems for
>
>sh> which all specifics of the physical implementation are irrelevant,
>
>sh> because every implementation of the same syntax is equivalent
>
>sh> in some respect (and the respect under scrutiny here is thinking).

When you describe computers as, "a subset of these analog systems [that] are what they are, and do what they do, purely because they are implementations of the right symbol system (because they are constrained by a certain formal syntax, manipulating discrete symbols on the basis of their arbitrary shapes: ...)", you are characterizing them at a level of description above that used to distinguish between the three causal mechanisms I claim to be important to this discussion. In the above quoted passage, all physical systems are classified as "analog", with a particular subset of these functioning in a specific way because "they are implementations of the right symbol system".

You go on to say that this functioning is the result of "manipulating discrete symbols on the basis of their arbitrary shapes" and then acknowledge that I refer to this as "pattern matching". But terms such as "manipulation" and "symbols" conform with the use of "pattern matching" as a *functional* description of a particular process, not a *physical* description of how that process physically accomplishes what it does. I believe this is, in part, one reason why (though it may just be a symptom) you distinguish only two kinds of things:

>
>sh> So I repeat, there seem to be TWO kinds of things distinguished
>
>sh> here (actually, one kind, plus a special subset of it), namely, all
>
>sh> physical systems, and then the subset of them that implement the
>
>sh> same syntax, and are equivalent in that respect, independent of the
>
>sh> physical properties of and differences among all their possible
>
>sh> implementations.

Again, "causal mechanism" is *below* the level of "symbol", which is an interpretive notion. Whether something is a symbol in a systematically interpretable symbol system is of no consequence to the physical process of pattern matching. What matters to pattern matching, as I use it, is physical structure (regardless of its physical realization -- electrical, biomolecular, etc.) and the structural "fitting" of physical structures. One can talk about pattern and matcher structures and an action triggered by a successful match (e.g., a particular voltage change or a single covalent bond), without ever invoking interpretive terminology such as "symbol" and "syntax".

Still at the physical level, we can also say that the cause of the voltage change or covalent bond formation was the result of the "fitting" of the two physical structures. It is not necessary to go into why this is physically so, but Pattee [1986] discusses it as do I, though with a different, complementary explanation [Boyle, in preparation]. Suffice it to say that it does, and there is no need to talk in terms of symbol manipulation, etc., in order to explain why. So my claim is that this particular structure-fitting process, regardless of the fact that it involves analog processes (due to various manifestations of electrical forces -- free charge, molecular, atomic), is one kind of causal mechanism: it enables extended physical structures to be causal.

I claimed in one of my previous posts that computational systems, if we are to avoid a vacuous definition of computation, must have pattern matching as the causal mechanism underlying what are referred to as "computational regularities", that is, the many-to-one relationship between physical and computational states in so-called computational systems [Pylyshyn, 1984]. This physically-principled criterion avoids over-interpreting planetary systems, for example, as being computational. Furthermore, this particular causal mechanism conforms with the more abstract notion of a system being "constrained by a certain formal syntax, manipulating discrete symbols on the basis of their arbitrary shapes". Pattern matching is real, relying on structural constraints to physically distinguish it from the collision of two billiard balls, for example.

>
>sh> But the passage above seems to imply that there is a THIRD kind
>
>sh> of stuff, that the brain will turn out to be that, and that that's the
>
>sh> right stuff (which Frank calls "intrinsic capacity for reference").
>
>sh> ...
>
>sh> fb> ...
>
>sh> I guess "SPS" is this third kind of property, but I don't really
>
>sh> understand how it differs from an ordinary analog process.
>
>sh> ...
>
>sh>
>
>sh> fb> What other ways might physical objects cause change besides
>
>sh> fb> through their [arbitrary, syntactic] forms? There are, I claim,

>
>sh> fb> only two other ways: nomologically-determined change and
>
>sh> fb> structure-preserving superposition (SPS). The former refers
>
>sh> fb> to the kinds of changes that occur in "billiard-ball collisions".
>
>sh> fb> They involve changes in the values of measured attributes
>
>sh> fb> (properties whose values are numerical, such as momentum) of
>
>sh> fb> interacting objects according to their pre-collisional measured-
>
>sh> fb> attribute values in a physically lawful way (that is, according to
>
>sh> fb> physical laws).
>
>sh> fb> ...
>
>sh> fb> Like pattern matching (PM), [SPS] also involves extended
>
>sh> fb> structure, but in a fundamentally different way. Whereas PM
>
>sh> fb> involves the fitting of two structures, which by its very nature,
>
>sh> fb> leads only to a simple change such as the switching of a single
>
>sh> fb> voltage value from "high" to "low" (in digital computers), SPS
>
>sh> fb> incolves the actual *transmission* of structure, like a stone
>
>sh> fb> imprinting its structure in a piece of soft clay.

First, the use of "stuff" to describe SPS gives the impression that it is some new kind of substance or that it depends on properties peculiar to a particular medium, rather than being a causal mechanism that describes a particular way in which physical objects can causally affect each other.

My three causal mechanisms are intended to describe all the ways in which physical objects can affect each other physically. That there are three is based on the claim that physical objects have two, what I call, physical "aspects" (for lack of a better term -- I don’t like to use "property"). These are 1) their measured attributes -- numerically-valued quantities like momentum whose values are constrained by physical laws and situation-specific constraints, and 2) their extended physical structures. This I take to be self-evident. Any other aspects that we associate with physical objects have to be functional or relational aspects which are interpretive notions, and, therefore, like "symbol", are abstractions above the present physical level of analysis.

Now, the only ways physical objects (whose interactions are what cause change in the world) can effect changes are by changing one (or both) of the physical aspects of other physical objects (as well as their own) when they interact. Furthermore, one (or both) of their physical aspects is (are) responsible for the resulting changes.

I've already described one of the three causal mechanisms above, which I call "pattern matching". For this process, it is the extended structure of an object that leads to physical change. And what sort of physical change does it lead to? A change in the value of a measured attribute, such as the voltage value of a particular circuit in a computer. Why does it lead to this kind of change? Because there is another physical structure which has a similar (perhaps identical, though that is not necessary, or complementary) pattern -- an arrangement of contours, combination of voltage values, etc. -- so that they physically "fit". Now it doesn't matter how this fitting occurs, whether it is a sort of "all at once" fitting, as in enzyme catalysis, or whether it occurs over a longer period of time and is spread out over different locations. The important characteristic is that the resulting change, which is due to all the individual (local) analog processes that underlie this kind of object interaction (e.g., individual transistor switchings due to electric charges in and potentials across semiconducting materials, or the local molecular forces involved in the structural positioning of biomolecules), happened because the object's structure acted as a constraint, "channeling" the effects of all these local changes into the final outcome. The change is simple because it is brought about by STRUCTURE FITTING.

Now you say that there are "all physical systems, and then the subset of them that implement the same syntax". But this kind of statement, which refers to the physical as analog and then uses interpretive language like "syntax" will not buy you the kind of physical distinctions I'm making. I divide everything except this special subset into two kinds of causal mechanisms: what I call nomologically-determined change and structure- preserving superposition. They are analyzed at the same descriptive level -- the physical/causal level -- that pattern matching was. Your analog vs syntactic seems to be a mixing of levels, whereas my three causal mechanisms are not.

Continuing with the type of analysis applied to pattern matching above, nomologically-determined change involves changes in the measured attribute values of interacting objects, caused by those objects' interaction and constrained according to physical laws and situation-specific constraints. This is what happens in any object interaction, even ones that involve structure fitting. But in structure fitting, one of the final changes is due to the fitting of structures, while the rest are nomologically determined. Most physical interactions in the world result in nomologically determined changes *only*. Such interactions are exemplified by billiard ball collisions.

So far we have structure leading to measured attribute changes (pattern matching) and measured attributes leading to measured attribute changes (nomologically determined change). Now what about structure leading to structural changes? That is what occurs in SPS. A stone colliding with a piece of soft clay is a model of such a process. The stone's surface structure is effectively transmitted to the clay. Of course, measured attributes of the clay and stone are also changed as a result of the collision, but the surface structure change in the clay is due to the surface structure of the stone, not its particular momentum value, for example. The causal mechanism is different than pattern matching because pattern matching involves structure fitting. For the latter, the effect is a measured attribute (structureless) change, rather than a change in structure as in SPS. The reason I use "superposition" to refer to this third causal mechanism is because it involves the superposing of one object's structure onto another's.

So all of the above causal mechanisms are PHYSICAL processes. I have not talked about symbols, syntax, representation, reference or anything of that nature, so I assume you will agree that there has been no hermeneutics creeping in so far. (At this point, one might ask: What about measured attributes leading to structure changes? This would occur, for example when two objects collide with such force that one or both break into smaller pieces. But the structures of the newly formed surfaces are due to the particular material and how the energy was channeled through it, probably due to its internal structure (cleavage planes, for example), so that the structures of the pieces are not really determined by any sort of relationships between them and the values of the measured attributes of the original objects. In other words, you can't have structureless entities creating non-arbitrary (with respect to the values of those entities) structures. Thus, I have lumped these kinds of effects in with nomologically determined changes.)

>
\>sh> Leaving out the hermeneutics of "appearance" (which I think is a
>
\>sh> dangerous red herring), the above again simply seems to be
>
\>sh> distinguishing two kinds of analog processes, but this time with
>
\>sh> the distinction mediated by properties that are interpretable as
>
\>sh> "resembling" something rather than by formal syntactic properties
>
\>sh> that are interpretable as meaning something. So, enumerating, we
>
\>sh> have (1) the usual Newtonian kind of interaction, as between planets,
>
\>sh> then we have (2) a kind of structure-preserving "impact," leaving an
>
\>sh> effect that is somehow isomorphic with its cause (like an object and
>
\>sh> its photographic image?), and then finally we have (3) implementation-
>
\>sh> independent semantically interpretable syntactic interactions. But (2)
>
\>sh> just looks like an ordinary analog transformation, as in transduction,
>
\>sh> which I don't think is fundamentally different from (1). In particular,
>
\>sh> if we drop talk of "appearances" and "resemblances," whatever
>
\>sh> physical connection and isomorphism is involved in (2) is, unlike
>
\>sh> (3), not merely dependent on our interpretation, hence not
>
\>sh> "ungrounded" (which is why I make extensive use of this kind of
>
\>sh> analog process in my model for categorical perception).

Not distinguishing between (2) and (1) leads to a problem similar to the one we have with describing digital computers; a descriptive dualism that talks about symbols, syntax and function in order to describe their computational behavior, and changes in voltage and other measured attributes of their physical components in order to describe their physical behavior. The first uses terms that are ungrounded, while the second uses physically grounded terms. If you want to understand the computational behavior of computers in a physically principled way, then you must ground its (internal) computational behavior. This is done via pattern matching and the causality of extended physical structure enabled by such a process. We cannot be satisfied with the standard functionalist gloss that there are causal relationships between computational states. This doesn't really ground the computational behavior of computers. It merely acknowledges the fact that computers are physical systems.

A similar argument can be made for distinguishing between SPS and standard physical descriptions. Clearly, when extended structure effects a structural change in other structures, it involves changes in those structures' local attributes (e.g., voltage values, say, in a neural network or positions of particular points on a structure's surface, say, in the stone/clay model). But describing such a structural process in this way -- e.g., as the transduction of photon energy to neuronal electric potentials -- loses the fact that there was a coherent structure which was the cause of the particular arrangement of individual local attribute changes and that this newly created structure may then go on to affect other structures. As with computation, it reduces the process to a set of changes described by standard physical state description terminology, so that if we wanted to consider that this kind of change is what underlies thinking (analogous to computation in the computer), then we would have to resort to information processing terminology in order to talk about it, like we do for computation; a terminology that is ungrounded.

Why do we need a physical framework based on the causal mechanisms I am proposing as responsible for physical change? Because the causality of extended structure is not explicitly accounted for in standard physical state descriptions. That is, for both computation and its counterpart in the brain -- thinking, however it is enabled -- it is the causality of extended structure,through pattern matching and SPS, respectively, that makes it computation and thinking, respectively. Just saying that there are analog processes is not sufficient, because all physical processes involve analog processes, including pattern matching (as you've acknowledged). Structure has to be recognized as controlling the behaviors of certain physical systems AS STRUCTURE, not just as a set of boundary conditions restricting the range of values of certain state variables. By lumping everything into analog processes or transformations that are non-syntactic, you are unable to distinguish between these.

Why should such a distinction matter? Because I believe the brain qua mind works at the level of structure transmission, no matter how much others want to reduce its behavior to neurophysiological descriptions based on measured- attribute analog transformations or transductions. If you don't ground structure transmission in some physical framework, then mind will always be described in ungrounded terms just as computation, described in terms of symbols and rules to match them, is ungrounded. This often leads people to believe that mind is emergent, rather than the result of a specific type of structural control.

I'm surprised that you don't see the necessity of this division, since it seems to me you would be hard pressed to explain how what you call "analog reduction" could produce iconic category structures if such a process were nothing more than lots of transductions of measured attributes of neurons without some more overarching structural constraints. Perhaps you don't think such constraints are necessary, but if that's the case, then all you can probably hope for is mind as an emergent property (which I don't agree with).

Finally, my use of "appearance" to describe extended structure is meant to distinguish how extended structure is causal in SPS, as opposed to how it is causal in pattern matching. For the latter, I call extended structure "form" because to be causal it must conFORM to another (matching) structure. Its structural *appearance* is not part of the effect. There is no interpretation involved in saying it this way; no smuggling in of a homunculus to figure out what the structure appears to look like (e.g., a tree, an elephant, etc.). "Appearance" and "form" are terms that are meant simply to help describe, in a more concise way, differences in how extended structure can effect change. It is still all physical, so I think I've steered clear of your "hermeneutical hall of mirrors".

>
>sh> My own proposal is that symbols are grounded in whatever
>
>sh> internal structures and processes are required to generate TTT
>
>sh> capacity, and I have no reason to believe that these consist of
>
>sh> anything more than (1) pure analog properties, as in solar
>
>sh> systems and their analogs, plus (2) syntactic properties, but
>
>sh> with the latter grounded in the former, unlike in a pure
>
>sh> (implemented but ungrounded) symbol system such as a
>
>sh> computer. In this hybrid system (Harnad 1992 -- see excerpt
>
>sh> below) neural nets are used to detect the invariants in the analog
>
>sh> sensory projection that allow object categories to be connected
>
>sh> the symbols that name them; this model invokes no third, new
>
>sh> property, just analog and syntactic properties.

Since I've already stated above why I believe there should be a subdivision of processes based on the causal mechanisms I've described here, as well as in previous posts and the literature, let me just comment briefly on your idea of a hybrid system. I think that the brain involves SPS (or in your terminlogy, is analog) "all the way through". Though there may be some pattern matching (or in your terminology, syntactic properties), I think this occurs at relatively "low level" perceptual stages, definitely not at higher cognitive levels. If, in your system, you "connect" object categories with the symbols that name them, AND the manipulation of those symbols according to their "syntactic properties" are what you intend to underlie thinking, then all you've really got is a pattern matching system with some peripheral grounding which I don't see as being different, in principle, than "a pure (implemented but ungrounded) symbol system such as a computer" for two reasons: 1) the symbols are still form-arbitrary because the connectionist network used to ground them is really just a pattern matching structure [Boyle, 1991], at least the way I've seen it described in your publications, and 2) even if the network is not a pattern matching structure (we could even assume that the iconic category structures are the symbols), the fact that the symbols are part of a "symbolic component" (i.e., they cause change through pattern matching) means that their referential capacities cannot be due to their structures since pattern matching renders *any* structure inconsequential with respect to the change it produces. Thus it wouldn't matter that they were "connected to" transduced category structures. In other words, the causal effects of representing entities are important to their referential capacities and, thus, to how they mean [Boyle, 1992], just as they are important to grounding. So if you want to be fundamentally different than a computer, the physical changes that underlie thinking cannot be due to pattern matching. Grounding is superfluous if it doesn't go "all the way in".

>
>sh> fb> Issues about consciousness, qualia, etc. should be part of
>
>sh> fb> another discussion on mind and brain, but symbol grounding
>
>sh> fb> and even the Chinese Room... should be part of the "What
>
>sh> fb> is Computation?" discussion because they involve issues of
>
>sh> fb> causality and representation which are fundamental to
>
>sh> fb> computation... e.g., "understanding" ...comes about
>
>sh> fb> presumably because of referential characteristics of the
>
>sh> fb> representation.
>
>sh>
>
>sh> But by my lights you can't partition the topic this way, excluding
>
>sh> the question of consciousness, because consciousness already
>
>sh> enters as a NEGATIVE datum even in the Chinese Room: Searle
>
>sh> testifies that he does NOT understand Chinese, therefore the
>
>sh> implementation fails to capture intrinsic reference. Searle is
>
>sh> reporting the ABSENCE of understanding here; that is an
>
>sh> experiential matter. So understanding piggy-backs on the capacity
>
>sh> to have qualia. Frank seems to agree (and to contradict this
>
>sh> partitioning) when he writes:

>
>sh>
>
>sh> fb> This is just a physical explanation of why, as Harnad puts it,
>
>sh> fb> there is "nobody home" in such systems. Nor can there ever be.

What I meant by the above is that discussions about consciousness and qualia, if they are not physically grounded (which includes most of the literature on these topics), should not be part of the discussion, "What is Computation?". But symbol grounding, reference and the Chinese room -- to the extent that it can be used to illustrate the arbitrariness of formal symbol systems -- are all relevant because they can be discussed in objective terms that are grounded. I don't know how to ground consciousness and qualia because I don't really know what kinds of things they are -- my current take is that they are epiphenomenal; the sensations we experience are the result of the particular causal mechanism underlying thinking and the particular large-scale organization of the brain. Understanding is sort of an in-between term which I believe can piggy back on reference.

Thus, if you can't find a common (physical) basis for aspects of cognition and for computation, then those aspects of cognition shouldn't be a part of the discussion. My reference to your "nobody home" characterization of computers was meant to acknowledge that at bottom one needs the right physical characteristics (for me, SPS), since they are, fundamentally, what give rise to our human brand of understanding and experience.

>
>sh> I know that "extended structure" plays a critical role in Frank's own
>
>sh> theory, but I have not yet been able to understand clearly what that role
>
>sh> is. Whenever I have read about it, if I subtracted out the hermeneutics, I
>
>sh> found no remarkable property left over -- other than continuity in time
>
>sh> and space, which is rather too general to be of any help, plus ordinary
>
>sh> analog and syntactic interactions.

If there are hermeneutical aspects to what I've outlined at the beginning of this post with respect to my causal mechanisms, please let me know what they are. It is true that I used the term "information" a bit loosely in the JETAI paper, but it was employed mainly for descriptive purposes, just as "appearance" is above. Nowhere, as far as I can tell, has the theory depended on such terms; there have been no mind-begging appeals to them.

I guess my response to "no remarkable property left over -- other than continuity in time and space, which is rather too general to be of any help" is: How do you see your distinction between pure analog and syntactic processes as being any less general?

Let me just close by giving a brief overview of what I see as the reason for our differences (because, except for the issue of grounding all the way in, I don't see our ideas as being that different). It seems that you are not making a distinction that is finer-grained than analog vs syntactic processes because you are relying on a variant of the Turing Test, your TTT, to test your grounding proposal. But I think one can go further than this in establishing what may be necessary for minds without having to go through the arduous process of building such a system (if, in fact, it is practical or even possible for us to do), perhaps only to find out that it can't be done (how do we know it can't and why if we never successfully build it) because it wasn't pure analog all the way in (though we might never know that). If we can ground thinking in the physical structures of the signals that project environmental structure (e.g., the structures of objects) onto our sensory surfaces, that is, determine the causal mechanism for how those signal structures alter the structure of the brain and, hence, influence its behavior, then we have a better chance of building a system that could pass any behavioral test we decide to throw at it. This is why I think we have to recognize a finer grained distinction within your analog processes category.

Well, I hope all this helped to clarify the points you raised. If not, perhaps the above will lead to other, more detailed comments and questions.

-Franklin Boyle

Boyle, C. F. (1991) On the Physical Limitations of Pattern Matching. Journal of Experimental and Theoretical Artificial Intelligence, 3:191-218.

Boyle, C.F. (1992) Projected Meaning, Grounded Meaning and Intrinsic Meaning. To appear in the Proceedings of the 14th Annual Conference of the Cognitive Science Society. To be given as a poster.

Boyle, C. F. (in preparation) The Ontological Status of Mental Objects.

Harnad, S. (1990) The Symbol Grounding Problem, Physica D, 42: 335-346.

Pattee, H.H. (1986) Universal Principles of Language and Measurement Functions In J.L. Casti and A. Karlqvist (eds), Complexity, Language and Life: Mathematical Approaches, (Springer-Verlag, New York)

Pylyshyn, Z. (1984) Computation and Cognition: Toward a Foundation for Cognitive Science, (MIT Press, Cambridge, MA).

Date: Thu, 2 Jul 92 14:43:37 EDT From: "Stevan Harnad" To: jfetzer@ub.d.umn.edu Subject: Publishing the "What is Computation?" Symposium

> From: jfetzer@ub.d.umn.edu (james fetzer)
> Subject: SPECIAL ISSUE OF MINDS AND MACHINES
> To: harnad@Princeton.EDU (Stevan Harnad)
> Date: Mon, 29 Jun 92 14:48:17 CDT
>
> Stevan,
>
> I wanted to contact you concerning our tentative plan for a special
> issue of MINDS AND MACHINES devoted to the topic, "What is
> Computation?" I have noticed the tremendous interest in this subject

> since the email exchange began as well as a considerable variation in
> opinion about how to go about pursuing the idea of publication. Based
> on my reading of the participants' reactions, my inference is that
> there is a strong preference for position papers, perhaps supplemented
> by critical discussions of one another's positions. That is an
> appropriate approach, it seems to me, where each participant can be
> assured of having their views presented intact in the form of a
> position paper, where opportunities for critical exchange are also
> provided (more on the order of the email origins, but now in relation
> to these more carefully considered position papers rather than the
> original formulations advanced earlier by email).

Jim, I have no objection to this alternative, if that is what the contributors prefer, but I have to point out that there is an element of wishful thinking in your reading of the participants' reactions as expressing "a strong preference for position papers." The latest tally of the votes had in fact been: 2 preferring position papers, 5 preferring the interactive symposium, and 11 amenable to either or a combination. Since then the tally has risen to 2/8/11, respectively, with all but 4 of the contributors now having cast their votes (see end of this message).

> I also believe that it is better to focus on the question of the nature
> of computation instead of combining this issue with questions about the
> symbol grounding problem. If you or others are inclined to contend that
> the one cannot be resolved without an adequate answer to the other,
> that of course would be an appropriate position to argue in your own
> position paper, but I think it would be an imposition to require others
> to focus on both when they may think that they are separable problems.
> (We do not want to beg the question by assuming they are wrong in
> advance.)

The ancillary issue is not the symbol grounding problem but whether or not cognition is a form of computation. That, after all, is what is motivating the "What is Computation?" question for most of us (and for your journal, "Minds and Machines," too, I should think). I happen to be among those who think that the "What is Computation?" question can and should be settled completely INDEPENDENTLY of the "Is Cognition Computation?" question, but I certainly would not force anyone to consider only one question or both, unless they feel so inclined. However, you will find, I think, that cognitive issues (including Searlian, anti-Searlean and symbol-grounding-related ones) will surface in the discussion whether the publication conforms more to the interactive symposium that is now transpiring (where cognitive issues are clearly being raised) or consists instead of position papers and subsequent interactive discussion, and whether or not its exclusive theme is "What is Computation?"

> If we agree that the combination of position papers and discussions
> is the right way to go, then let me suggest that we target this special
> issue to be the November 1994 issue of this journal. I would like to
> have everything in my hands no later than April 1994, where you need to
> have everything in your hands much sooner to make it all come out
> right. This should provide sufficient time for the contributors to
> compose their papers and exchange them prior to creating the critical
> exchange parts of the issue. I would stress that I believe that this

> should be done in a certain sequence to preserve the integrity of
> various authors' positions. I imagine you will want to continue your
> ongoing exchanges via email as a separate undertaking even while this
> project develops for MINDS AND MACHINES.

I will continue the email symposium as long as the ideas keep flowing. In September I will ask the contributors whether they wish to prepare formal position papers for a further round of email discussion with a view to publication. A tentative target date for receiving and circulating the position papers electronically might be November, when they could be refereed and edited, say, by January. Then the accepted versions could be circulated for electronic commentary and cloture on the ensuing discussion might be invoked in May or June, when the discussion can be edited, returned to the contributors for approval or modification, refereed, re-edited, and then sent to press.

> Let me know if this sounds all right to you. The number of words per
> pages of this journal is 400 rather than 600 (as I believe I mistakenly
> indicated previously). I am willing to commit 100 pages to this
> undertaking and possibly more if it turns out to warrant a greater
> commitment. We will run 125 pages per issue beginning with volume 3
> (1993), but I would like to keep 25 pages for book reviews and such,
> even in the case of special issues, if it is possible. Given what I can
> do, I will keep the reviews on related issues. So let me know if this
> sounds agreeable and how you plan to proceed, etc., and we can carry
> this project forward within these parameters. I am very enthusiastic
> about it.
>
> Best wishes, Jim
>
> ase note the sarcasm dripping off the word "benefit" above.

Those parameters seem ok to me. Since I happen to favor the interactive symposium format, I would prefer short position papers and extended discussion, rather than extended position papers and short discussion, but I leave that to the contributors to decide collectively.

> From: jfetzer@ub.d.umn.edu (james fetzer)
> Date: Wed, 1 Jul 92 11:00:41 CDT
>
> Stevan,
>
> Let me know if the general outline I have sketched sounds agreeable.
> Perhaps it might be possible to have it in print sooner than the
> November 94 issue (as I am sure you would like). More traditional
> preparation, as you know, requires long lead times, such as about two
> years for special issues. So this case--where the preparation of
> manuscripts might take place in much less time--poses special
> circumstances. Let me know how you feel about the timing. I would
> have no objection to the idea of your making this material available
> via email earlier than its publication in MINDS AND MACHINES. That
> could compensate for the timing, although I do not know the time frame
> you have in mind.

>
> Jim

Let’s see how the participants feel about the tentataive dates and formats. Below is the lastest vote tally.

Cheers, Stevan

Previous tally of votes:

Interactive Symposium (IS) vs. Position Papers (PP): Either or Combination: 11 - Prefer IS: 5 - Prefer PP: 2

With the further votes below, the tally is now 11/8/2

Date: Wed, 20 May 92 22:59:40 EDT From: lammens@cs.Buffalo.EDU (Joe Lammens)

I’m for publication, preferably as an interactive symposium, perhaps with position papers added.

Date: Mon, 25 May 92 00:39:00 +0100 From: chrisley@csli.stanford.edu Ronald L. Chrisley

Publishing is fine by me. I have no strong preference concerning the format.

I’ve found the discussion very interesting, and have some points I’d like to make, but I have not been able to find time to catch up with the postings. I hope I will find time soon.

Could you give me a run-down of the agenda? When will discussion end? When will position papers be expected?

Date: Thu, 4 Jun 1992 10:49:33 PDT From: Patrick Hayes

[Voted earlier, but here clarified his vote, so re-assigned to IS]

Sorry Im late, hope not too late. I like to try the edited interactive format, although I think it will require a masterpiece of editing to get it sorted out and readable. I sympathise with some of Brian Smith’s concerns also, and of course the ideal situation would be an editor who had no position of his own, but thats probably impossible to achieve here. This is not like a taperecording of a face-to-face conversation, but in any case that can be made readable, with enough work and some ruthlessness in cutting inappopropriate chunter. (Ive done it to transcripts of arguments between groups of cognitive scientists.)

In fact its not like anything else, which is why I like the idea of trying to make it into something. I share Stevan's (or should I say 'your': am I talking to Stevan or the CC list? One of those interesting email questions) fascination with the evolution of a new medium, and would like to help xperient with it.

Pat Hayes

Date: Tue, 16 Jun 92 16:44:27 -0400 From: hatfield@linc.cis.upenn.edu (Gary Hatfield)

If you publish some chunk of the interactive discussion and want to include my part, I would give permission with the condition that I be allowed to review and approve the material.

From: massimo@Athena.MIT.EDU Massimo Piatelli-Palmarini Date: Wed, 24 Jun 92 08:54:10 EDT

[From a reader of the discussion, not yet a contributor]

Dear Stevan, I am back from a long absence and wish to say that the idea of publishing the debate is a good one, provided that a lot, I mean a lot, of editing is carried out on the existing exchanges.

Date: Fri, 3 Jul 92 12:06:53 EDT From: "Stevan Harnad" To: jfetzer@ub.d.umn.edu Subject: Re: Publishing the "What is Computation?" Symposium

```
> From: jfetzer@ub.d.umn.edu (james fetzer)
> Date: Thu, 2 Jul 92 14:59:21 CDT
>
> Stevan,
>
> What you have in mind is fine. Longer discussions and shorter position
> papers is probably the right combination: my suggestion was meant to
> emphasize the desirability of having position papers to make clear (in
> condensed form) the positions of various contributors. Not all of them
> need to have papers rather than comments, of course. But it seems to me
> that having to read through 100-pages of interactive discussion
> WITHOUT POSITION PAPERS would not be something we should impose on our
> readers. If you can provide 100 pages at 400 words per page of material
> MAX by April 1994--watch out for spacing, which must be taken into
> account--I can schedule it for the November 1994 issue now and we are
> on track.
>
> To clarify the dates mentioned in my last message, I need the final
> version in hand by April 1995. But we also need to have the referee
> process take place, as indeed your tentative schedule clearly accommo-
> dates. So everything sounds fine to me.
>
```

> Jim

Except that I think you do mean April 1994, n'est ce pas, unless you're REALLY pessimistic about the refereeing...

Chrs, Stevan

Date: Fri, 3 Jul 92 14:34:21 EDT From: "Stevan Harnad" Subject: Re: What is Computation?

Date: Fri, 3 Jul 1992 12:17:42 -0400 (EDT) From: Franklin Boyle

David Chalmers writes:

>dc> Thanks to Franklin Boyle for his thoughful replies. I confess to not
>dc> fully understand his position, but as far as I understand it, I gather
>dc> he's saying that for the purposes of determining what's computation
>dc> and cognition, we have to look more closely than simply at the causal
>dc> state-transitional structure. What matters isn't just the pattern of
>dc> transitions, it's (a) the specific nature of the state-transitions,
>dc> and (b) the specific nature of the causal relations.
>dc>
>dc> As far as (a) is concerned, we have to distinguish between real
>dc> changes in structure, e.g. "receiving a projection of a structure",
>dc> from mere changes in the "measured attributes" of a system (e.g.
>dc> voltages). As far as (b) is concerned, we have to distinguish between
>dc> "structure-preserving superposition", in which the form (or
>dc> appearance) of one state somehow imprints itself on another, from
>dc> mere "pattern matching" and "structure fitting". The wrong kinds of
>dc> state-transition and causation give you computation; the right kinds
>dc> might give you cognition.
>dc>
>dc> My reply to this proposal is pretty simple: I'm not sure that these
>dc> distinctions come to anything, and I see no reason why they
>dc> should make a difference between cognition and non-cognition.
>dc> It seems to me that the form or appearance that various states
>dc> embody is just irrelevant to a system's status as cognitive, or
>dc> as computational.
>dc>......
>dc> I've probably misunderstood this position completely,....

I would say you gave a pretty fair, albeit brief, summary of the main ideas.

The reason I believe the "form or appearance" of a system's states (actually of its constituent structures such as neural connectivity and the activation patterns constrained by it) are relevant to a system's status as cognitive is because in order to "know" about objects, events, etc. in the world, we have to somehow acquire information about their structures. If we don't get this spatially varying structure projected directly, then it implies that some kind of external agency set up an apparatus to analyze the structures into whatever *encoded* form is desirable.

This is what happens in digital computers. We encode descriptions (say propositional) of the environment and then have the computer associate those descriptions with actions, say, through rules whose left-hand sides match the descriptions. This is accomplished by the physical process of pattern matching, so that as long as matchers are available for triggering the appropriate actions, it doesn't matter what the encodings look like (they could even be bitmaps of the environment). The system, say a robot with the full range of human peripheral capacities, would simply need a decoder hooked up to its effectors in order to be able to behave in a manner consistent with what we interpret the descriptions to be about.

So, I don't see how we can be cognizant of structures in the environment if we don't have information in our heads which was constructed from those structures such that the structural information is *actually* there. And the structural information isn't there if it's encoded. We only think it is because, as outside observers, we can *interpret* what is there as such and observe behaviors which are consistent with that interpretation. Why would a propositional representation be any different than arbitrary bit strings, neither of which preserves the spatial variations in any form which is, spatially, even remotely isomorphic to the external structures they purportedly represent? In a pattern matching system, these arbitrary forms can represent whatever we want them to. To repeat, if the structural variations are not actually topographically preserved, (and there appear to be many topographic mappings in the brain, at least in the sensory cortical areas) then how can we be said to have information about those structures except by an observer looking into our heads (or into a computational model) and interpreting the representing entities to be such?

Certainly there are physical changes to the visual signal entering the head on its way to the visual cortex and beyond. But I believe these are enhancements and analyses which are accomplished within the spatially-preserved variations of the input signal.

In short, the right causal structure (sequences, etc.) is necessary for a system to behave *as if* it is cognitive. But structure preservation and its causal capacity for effecting change through superposition are also necessary for there to actually be cognition; that is, for having the capacity to, for example, "understand" in Searle's sense.

>dc> fb> If we allow any physical system to be an implementation
>dc> fb> of some computation, we will most likely end up with
>dc> fb> little in the way of principled criteria for determining
>dc> fb> whether cognition is computation.
>dc>
>dc> Let's dispose of this canard for once and for all. Even if
>dc> every system implements some computation, this doesn't
>dc> imply that every system is engaged in cognition, for the
>dc> simple reason that only *certain kinds* of computation
>dc> qualify as cognition. Not even the strongest of believers in
>dc> in strong AI has said that implementing *any* program is
>dc> sufficient for cognition. It has to be the right kind of program
>dc> (or, more generally, the right kind of computation).

Maybe I'm misinterpreting the expression, "cognition is computation", but it seems like you're interpreting what I said above to mean "computation is cognition", which I don't believe at all and did not intend for the above to mean.

>dc> Isolating those kinds of computation that qualify as cognition is
>dc> an interesting, highly non-trivial question in its own right.
>dc> Presumably, only certain highly complex computations will
>dc> be sufficient for cognition; solar systems, along with rocks and
>dc> most everything else in the world, won't have the requisite causal
>dc> structure to qualify.

I certainly agree that a certain level of complexity is necessary.

>dc> There's some kind of strange, deeply-embedded assumption here:
>dc> that true "knowledge" requires embedding of an object's actual
>dc> "structure" inside a cognitive system. To think about spheres, does
>dc> one need something spherical inside one's head? Surely not. This
>dc> sounds like the kind of scholastic theory that Cummins dismisses
>dc> in the first few pages of his book.

As Cummins points out, this kind of literal representation only works for mind-stuff considered as non-physical: "The idea that we could get redness and sphericity in the mind loses its plausibility if this means we have to get it in the brain. When I look at a red ball, a red sphere doesn't appear in my brain". (Cummins, 1989 p31). This is also Roger Shepard's "first-order isomorphism" (see Palmer, 1978). (I have not read Shepard's paper, but I list the reference below as cited by Palmer).

No, I don't take it quite so literally. I guess my ideas are more akin to what Cummins calls "restricted similarity" in the sense that spatial variations of the input are preserved, but that doesn't mean such a representation shares any other properties with its referent. It's just that the spatial variations of one can be mapped onto the other, though there may be metric deformation of the representing form.

>dc> Even if you don't mean something quite as literal as this, there
>dc> seems to be nothing wrong in saying that the brain embodies the
>dc> information it carries in mere "measured attributes" such as
>dc> potentials, frequencies, and so on, as long as these bear the
>dc> requisite causal relation to the outside world and play the
>dc> appropriate functional role within the system.

There are two things to be sorted out here. It is certainly the case that how the structural variations of the image on the retina, for example, get topographically mapped onto the visual cortex may occur by transducing light energy amplitudes into frequencies of spike trains along axons, so in this sense there are measured attributes which "carry" the "information" of the signal. But I claim the significant behavior of the brain depends on the *spatial variation* of these measured-attribute values which cause the structural variations to become "imprinted" (temporarily) on the visual cortex. A particular measured attribute value, or a set of values that are not spatially arranged such that variations in their values are isomorphic (modulo metric deformations) to the variations of the original signal, cannot be said to carry structural information about the referent unless we interpret it, from the outside, to be such.

-Franklin Boyle

Cummins, R. (1989) Meaning and mental representation (MIT Press/Bradford Book).

Palmer, S.E. (1978) Fundamental aspects of cognitive representation. In Rosch, E. and Lloyd, B. (eds), Cognition and Categorization, (Hillsdale, NJ: Lawrence Earlbaum)

Shepard, R. and Chipman, S. (1970) Second-order isomorphism of internal representations: Shapes and states. Cognitive Psychology 1: 1-17.

Date: Mon, 23 Nov 92 23:39:12 EST From: "Stevan Harnad" To: harnad@rrmone.cnrs-mrs.fr Subject: New Symbol Grounding Discussion

To: Symbol Grounding Discussion group

Well it's November 1992 and the "What Is Computation?" position papers are either just about to be submitted or already submitted and headed for refereeing, so here is a new topic. Selmer Bringsjord, in his recent book, "What Robots Can and Cannot Do" has proposed some arguments against the possibility that certain mechanisms can have minds. One of these, developed more fully in a paper of his, is an "Argument from Serendipity" against the Turing Test. I have asked him to reproduce this argument here, in what started as a one-on-one discussion, but we have both agreed that it's time to let the group join in. (Selmer, a late-comer to the group, is also at work catching up on the "What Is Computation" archive so he can submit a position paper too. Here are the first iterations (under 500 lines).

The next posting after this one (not the one below, but the next one under the header "TT and Necessity") will be open for discussion. You can of course quote from and comment on the exchange below as well.

Stevan Harnad

Date: Sat, 7 Nov 92 01:55:44 EST From: "Stevan Harnad" To: brings@rpi.edu (Selmer Bringsjord) Subject: Re: TTT

Selmer,

Can you generate a screen-readable ascii version of your serendipity argument against TTT? Or, better still, can you tell it to me in a nut-shell? I pride myself -- or candidly confess -- that I have never had (nor heard) an argument (other than a formal mathematical proof) that I could not tell to a motivated interlocutor in 5 minutes or 1500 words. Whenever I have an idea that takes more words, I suspect the idea.

Stevan

P.S. Let me guess that your serendipity argument is that something could pass the TTT by chance, just happening to have anticipated a lifetime of verbal and robotic contigencies correctly. My answer is that not only is such a combinatorial outcome surely NP-complete (and hence not really worth worrying about) but, as I insist below, Turing Testing is not a game: We really want to design a mechanism with TTT power, capable of handling the infinity of contingencies an infinity of lifetimes (or our own, if we were immortal) could encounter (even though it can only be TESTED in one finite lifetime). And if that's not enough to convince you that serendipity is no refutation of the TTT, let me remind you that we are not talking about necessary and sufficient conditions here: A non-TTT passer, indeed a rock, could, logically speaking, be conscious, and a TTT-passer could fail to be conscious. So, since logical possibility is not the issue in the first place, the monkey-at-a-typewriter combinatorial possibility is no challenge to a principled generative model of Shakespeare.

From: Selmer Bringsjord Date: Mon, 9 Nov 92 13:26:00 -0500 To: harnad@Princeton.EDU Subject: TT, TTT

Hello Stevan,

On TTT:

We agree that we *want* to design a mechanism with TTT power. (Indeed, prolly both of us are more than willing to work toward designing such a mechanism.) We agree that we (humans) *can* design a mechanism w/ TTT power. We also agree that it's logically possible that a TTT-passer could fail to be conscious. What dispute remains? Well, the only remaining issue for me (for the most part, anyway) is the central Turingish conditional; I think that it's false, and that's all I hope to show by the argument from serendipity. If there's a disagreement between us, you must think the conditional is true. But what is the conditional, and *is* is true?

The simplest construal of Turing's conditional (sparked by Professor Jefferson's Lister Oration in the original paper) is

(TT-P) Ax(x passes TT -> x is conscious)

where -> is the material conditional. But (TT-P) cannot be what is intended, since on standard model-theoretic semantics for FOL (TT-P) is vacuously true. On the other hand, neither can this conditional work:

(TT-P') L[Ax(x passes TT -> x is conscious)]

Because this proposition is overthrown (in standard modal contexts) by what you concede, viz.,

(1) M[a passes TT & ~a is conscious]

What construal remains? In my paper I respond to those who say that though Turing proposed something like (TT- P'), what he *should* have championed was a subjunctive or probabilistic conditional. The best response, which I don't make in the paper, is simply to demand a formal account of such conditionals (because absent such an account I'm well within my rights in refusing to affirm the conditional).

Yours, Selmer Bringsjord Dept. of Philosophy selmer@rpi.edu Dept. of Comp. Sci. selmer@rpitsmts RPI F: 518-276-4871 Troy NY 12180 USA P: 518-276-8105

Date: Mon, 9 Nov 92 16:20:39 EST From: "Stevan Harnad" To: brings@rpi.edu Subject: Re: TT, TTT

Hi Selmer,

We have a basic methodological disagreement, but it's not specifically about the TT. It's about the value of formalization on topics like this one. I find it useful in technical areas such as mathematics and computer programming, but I do not find it at all helpful in areas where things can be discussed in plain English. I am not a philosopher, yet I can speak with full understanding and authority about the premise that a candidate who passes the TT or TTT (which is just shorthand for exhibiting verbal or verbal+robotic capacities that are indistinguishable from our own) is or is not conscious. It is obvious that no one has a proof that either candidate MUST be conscious, hence it follows that neither is necessarily conscious.

Now, can you please reformulate in English the substance of what is at issue over and above this very simple and straightforward point on which I do not disagree?

Best wishes, Stevan

From: Selmer Bringsjord Date: Wed, 11 Nov 92 14:56:24 -0500 To: harnad@Princeton.EDU Subject: TT...

Okay, this may be somewhat better:

We have the methodological difference to which you allude in your previous note. And we have a difference, perhaps, over the conditional at the heart of Turing's case for TT. Let me try now not only to address the second, but the first also - in one stroke.

The conditional, in general, is simply the main thesis which Turing advanced, and which many thinkers since have likewise promoted: if something x passes the TT, then x is conscious. It was never enough just to state the TT and pack up and go home; *Mind* wouldn't have accepted the paper in that case. Your TTT and TTTT are intrinsically interesting, but what excites people about them is that perhaps *they* can supplant TT in Turing's conditional to yield a true proposition! It's the same situation with Turingish tests which can't in principle be passed by finite state automata, or Turing machines: the tests themselves are interesting, but the key is that they are supposed to help Turing's case. (This point is made especially vivid by the fact that people have proposed tests which no physical artifact could pass - on the reasonable assumption that machines beyond TMs will be, to put it mildly, rather hard to build. Certainly such tests are devised only to produce a defensible conditional (they may not succeed), not to give rise to a concrete empirical goal toward which AI should strive. But both of us do view TT and the like to be in part an empirical goal worth shooting for.) At any rate, I have expressed the rough-and-ready conditional in English; it isn't formalized. In *general* such informality, coming at a crucial dialectical juncture, worries me; in *general* the informality is something you find welcome. But I think we have here a case where the informal is unfortunate, as the following reasoning may show.

The simplest construal of Turing's conditional (sparked by Professor Jefferson's Lister Oration in the original paper) is

(TT-P) For every x, if x passes TT, then x is conscious.

where the if-then here is the material conditional. But (TT-P) cannot be what is intended, since on standard model-theoretic semantics for first-order logic (TT-P) is vacuously true. Because (TT-P) then says that for every element of the domain, if it passes TT, then it is conscious - and the if-then here is again the material conditional. Since no element of the domain passes TT, the antecedent is always false, and therefore by the characteristic truth-table for the material conditional the conditional is true.

On the other hand, neither can this conditional work:

(TT-P') It's logically necessary that (TT-P).

The operator here I wrote before as 'L' for the necessity operator in modal logics, more often a box. Now (TT-P') is overthrown (in standard modal contexts, i.e., if the necessity and possibility operators have standard meanings cashed out by normal systems of modal logic) by what we both affirm, viz.,

(1) It's logically possible that some thing can pass TT but not be conscious.

The proof that (1) implies not-(TT-P') is elementary, and as you say, we don't want it in this context anyway.

So then what construal remains? How can we get Turing off the ground? In my paper I respond to those who say that though Turing proposed something like (TT- P'), what he *should* have championed was a subjunctive or probabilistic conditional. The second of these possibilities would be that the conditional wouldn't be material in form, but something like P probabilistically entails Q. Enormous work has gone into trying to say what such a conditional amounts to - but there's no consensus at all, and in fact it's all rather messy. So it's like getting Turing out of his mess at the cost of tossing him into quicksand. This is a hard-nosed attitude (I take a different tack in the paper), I know, but I suspect it's an attitude that, based on what you said in your previous note, you would find acceptable in programming contexts, etc. There's a slogan "Theorems before programs." I like "Formal arguments about theorems before programs".

The subjunctive approach, at least the simplest version of it, is to augment (TT-P) with some such thing as: TT must be such that had it been run in other circumstances at other times, it would *also* be passed. Here again, though, the question of counterfactuals has occupied and continues to occupy logicians and philosophers of science and language - and there's no consensus. And anyway I try to show in the paper that on the dominant view of counterfactuals Turing's subjunctivized conditional is falsified by the argument from serendipity.

It seems to me that here we have a siuation which, with gem-like clarity, shows that the English (Turing's English) doesn't serve us well at all. It may be, however, that there is another construal (perhaps involving a 'going to be true' operator from temporal logic) which is defensible.

Everything I’ve said is of course consistent with my position that our robots will pass TT, TTT, ..., but will not be persons. In the present context, you could say that this is partly because persons are essentially (self-) conscious, and standarly conceived robots aren’t. (Chapter IX in WRC&CB is an argument for this.)

Right now we’re covering the symbol grounding problem in my course Philosophy of AI (upper undergraudate level), and my students greatly enjoy your writings on the matter. We may arrive at something worth sending on to you. Were looking now at whether some such proposal as "a formula (P) in some robot R’s KB means P for R iff some causal relation obtains between the R’s sensors and effectors, the external physical world, and KB" is promising. This is a proposal which Georges Rey has made, and it seems related to your proposal for how to ground symbols.

Yours, Selmer

Date: Sun, 22 Nov 92 22:30:35 EST From: "Stevan Harnad"

>sb> From: Selmer Bringsjord
>sb> Date: Wed, 11 Nov 92 14:56:24 -0500
>sb>
>sb> We have the methodological difference to which you allude in your
>sb> previous note. And we have a difference, perhaps, over the conditional
>sb> at the heart of Turing’s case for TT. Let me try now not only to
>sb> address the second, but the first also - in one stroke.
>sb>
>sb> The conditional, in general, is simply the main thesis which Turing
>sb> advanced, and which many thinkers since have likewise promoted: if
>sb> something x passes the TT, then x is conscious. It was never enough
>sb> just to state the TT and pack up and go home; *Mind* wouldn’t have
>sb> accepted the paper in that case. Your TTT and TTTT are intrinsically
>sb> interesting, but what excites people about them is that perhaps *they*
>sb> can supplant TT in Turing’s conditional to yield a true proposition!

Fine, but let me immediately add a clarification. I think it was arbitrary for Turing to formulate his criterion in the affirmative. The correct version would be: If a candidate passes the TT we are no more (or less) justified in denying that it has a mind than we are in the case of real people. That’s the interesting methodological thesis (false in the case of the TT, by my lights, but true in the case of the TTT, again by my lights, and overdetermined in the case of the TTTT) that I, at any rate, find worthy of empirical testing and logical analysis. Any stronger thesis just increases the quantity of arbitrariness in the problem, in my view.

>sb> It’s the same situation with Turingish tests which can’t in principle
>sb> be passed by finite state automata, or Turing machines: the tests
>sb> themselves are interesting, but the key is that they are supposed to
>sb> help Turing’s case. (This point is made especially vivid by the fact
>sb> that people have proposed tests which no physical artifact could pass -
>sb> on the reasonable assumption that machines beyond TMs will be, to put
>sb> it mildly, rather hard to build. Certainly such tests are devised only

>sb> to produce a defensible conditional (they may not succeed), not to give
>sb> rise to a concrete empirical goal toward which AI should strive.

Quite so; and I, for one, have no particular interest in defensible conditionals with no empirical content or consequences -- not on this topic, at any rate.

>sb> But both
>sb> of us do view TT and the like to be in part an empirical goal worth
>sb> shooting for.) At any rate, I have expressed the rough-and-ready
>sb> conditional in English; it isn't formalized. In *general* such
>sb> informality, coming at a crucial dialectical juncture, worries me; in
>sb> *general* the informality is something you find welcome. But I think we
>sb> have here a case where the informal is unfortunate, as the following
>sb> reasoning may show.
>sb>
>sb> The simplest construal of Turing's conditional (sparked by Professor
>sb> Jefferson's Lister Oration in the original paper) is
>sb>
>sb> (TT-P) For every x, if x passes TT, then x is conscious.
>sb>
>sb> where the if-then here is the material conditional. But (TT-P) cannot
>sb> be what is intended, since on standard model-theoretic semantics for
>sb> first-order logic (TT-P) is vacuously true. Because (TT-P) then says
>sb> that for every element of the domain, if it passes TT, then it is
>sb> conscious - and the if-then here is again the material conditional.
>sb> Since no element of the domain passes TT, the antecedent is always
>sb> false, and therefore by the characteristic truth-table for the material
>sb> conditional the conditional is true.

Still not English enough. I assume what you mean is that if anything OTHER THAN US passes the TT, then it's conscious -- and nothing other than us passes the TT, so the claim is trivially true. But if we construe this empirically, there may eventually be something other than us that passes the TT, and there are already coherent things we can say even about that hypothetical future contingency (e.g., Searle's Chinese Room Argument and my Symbol Grounding Problem).

>sb> On the other hand, neither can this conditional work:
>sb>
>sb> (TT-P') It's logically necessary that (TT-P).
>sb>
>sb> The operator here I wrote before as 'L' for the necessity operator in
>sb> modal logics, more often a box. Now (TT-P') is overthrown (in standard
>sb> modal contexts, i.e., if the necessity and possibility operators have
>sb> standard meanings cashed out by normal systems of modal logic) by what
>sb> we both affirm, viz.,
>sb>
>sb> (1) It's logically possible that some thing can pass TT but not be
>sb> conscious.
>sb>
>sb> The proof that (1) implies not-(TT-P') is elementary, and as you say,

>sb> we don’t want it in this context anyway.

Not only is it elementary, but no "proof" is necessary, because, as I said earlier, the affirmative thesis is too strong, indeed it’s an arbitrary claim. The only thing that would have given the TT the force of necessity would have been a PROOF that anything that passed it had to be conscious. No one has even given a hint of such a proof, so it’s obvious that the thesis is not stating a necessary truth. In its positive form, it is just an empirical hypothesis. In its negative form (as I stated it), it is just an epistemic or methodological observation.

>sb> So then what construal remains? How can we get Turing
>sb> off the ground? In my paper I respond to those who say that though
>sb> Turing proposed something like (TT- P’), what he *should* have
>sb> championed was a subjunctive or probabilistic conditional. The second
>sb> of these possibilities would be that the conditional wouldn’t be
>sb> material in form, but something like P probabilistically entails Q.
>sb> Enormous work has gone into trying to say what such a conditional
>sb> amounts to - but there’s no consensus at all, and in fact it’s all
>sb> rather messy. So it’s like getting Turing out of his mess at the cost
>sb> of tossing him into quicksand. This is a hard-nosed attitude (I take a
>sb> different tack in the paper), I know, but I suspect it’s an attitude
>sb> that, based on what you said in your previous note, you would find
>sb> acceptable in programming contexts, etc. There’s a slogan "Theorems
>sb> before programs." I like "Formal arguments about theorems before
>sb> programs".

It never even gets to that point. I don’t know why you even invoke the terminology of "subjunctive or probabilistic conditional": In its affirmative form it is an over-strong and probably untestable empirical hypothesis (which, like all empirical hypotheses, depends on future data that can be adduced for and against it) and in its proper negative form it is merely a methodological observation (perhaps also open to counterevidence or logical counter-examples, only I haven’t seen any successful instances of either yet). In my view, nothing substantive can come out of the formal analysis of the terms in which this simple thesis is stated.

>sb> The subjunctive approach, at least the simplest version of it, is to
>sb> augment (TT-P) with some such thing as: TT must be such that had it
>sb> been run in other circumstances at other times, it would *also* be
>sb> passed. Here again, though, the question of counterfactuals has
>sb> occupied and continues to occupy logicians and philosophers of science
>sb> and language - and there’s no consensus. And anyway I try to show in
>sb> the paper that on the dominant view of counterfactuals Turing’s
>sb> subjunctivized conditional is falsified by the argument from
>sb> serendipity.

Look, let’s make it simpler. Here’s a TT (actually a TTT) for an airplane: If it has performance capacity Turing-indistinguishable from that of an airplane, it flies. No subjunctives needed. Of course you can gerrymander the test of its flying capacity for a finite amount of time to make it appear as if it can fly, even though it really can’t; perhaps, with sufficient control of every other physical object and force for the rest of time (by God or by Chance -- "serendipity") you could do it forever. So what? We’re not interested in tricks, and the issue is not one of necessity or of subjunctives. There's a certain functional capacity a plane has, we call that flying, and anything else with the same functional capacity also flies. There is no higher authority.

With consciousness, however, there is a higher authority, and that concerns whether subjective states accompany the requisite performance capacity. Perhaps they do, perhaps they don't. My version of the TTT just makes the methodological point that we cannot hope to be the wiser, and hence it is arbitrary to ask for more of machines than we can expect of people. Again, no subjunctivities or necessities about it.

>sb> It seems to me that here we have a siuation which, with gem-like
>sb> clarity, shows that the English (Turing's English) doesn't serve us
>sb> well at all. It may be, however, that there is another construal
>sb> (perhaps involving a 'going to be true' operator from temporal logic)
>sb> which is defensible.

I don't think English is at fault; I think you are trying to squeeze necessity out of an empirical hypothesis, and I don't see any reason why you should expect to be any more successful here than with F = ma (which is not to say that TT claims are entirely like ordinary empirical hypotheses).

>sb> Everything I've said is of course consistent with my position that our
>sb> robots will pass TT, TTT, ..., but will not be persons. In the present
>sb> context, you could say that this is partly because persons are
>sb> essentially (self-) conscious, and standarly conceived robots aren't.
>sb> (Chapter IX in WRC&CB is an argument for this.)

I would not say anything of the sort. I think it is possible that all TTT-passers will be conscious and I think it is possible that not all TTT-passers will be conscious. I just think there are methodological reasons, peculiar to the mind/body problem, why we can never know one way or the other, even with the normal level of uncertainty of an ordinary empirical hypothesis. (By the way, although I believe that Searle's Chinese Room Argument has shown that it is highly improbable that TT-passing computers are conscious, it is not a PROOF that they cannot be; a second consciousness of which Searle is not conscious is still a logical possibility, but not one worthy of much credence, in my view.)

>sb> Right now we're covering the symbol grounding problem in my course
>sb> Philosophy of AI (upper undergraudate level), and my students greatly
>sb> enjoy your writings on the matter. We may arrive at something worth
>sb> sending on to you. Were looking now at whether some such proposal as "a
>sb> formula (P) in some robot R's KB means P for R iff some causal
>sb> relation obtains between the R's sensors and effectors, the external
>sb> physical world, and KB" is promising. This is a proposal which Georges
>sb> Rey has made, and it seems related to your proposal for how to ground
>sb> symbols.

As you know, I haven't much of a taste for such formalism. If what you/he are saying is that it is more probable that a symbol system is conscious if it can not only pass the TT, but also the TTT, that is indeed what I was saying too. And in particular, whereas a symbol in a TT-passing symbol system obeys only one set of constraints (the formal, syntactic ones that allow the system to bear the weight of a systematic semantic interepretation), a TTT-passing system obeys a second set of constraints, namely. that the system must be able (Turing-indistinguishably) to pick out (discriminate, manipulate, categorize, identify and describe) all the objects, events and states of affairs that its symbols are systematically interpretable as denoting, and these two different sets of constraints (symbolic and robotic) must square systematically and completely with one another. This would solve the symbol grounding problem -- the interpretations of the symbols in the system would be autonomously grounded in the system's robotic capacities rather than having to be mediated by the mind of an external interpreter) but it would still be a logical possibility that there was nobody home in the grounded system: no one in there for the symbols to mean anything TO. Hence groundedness and consciousness are certainly not necessarily equivalent. Indeed, I know of no way to prove that groundedness is either a necessary or a sufficient condition for consciousness. It's just a probable one.

Stevan Harnad

PS I think it's time to start posting this correspondence to the symbol grounding group as a whole. Confirm that this is OK with you and I will gather together the pertinent parts of what came before and post it, so the Group (about 500 strong) can join in. It would be a good way to start this year's discussion, which has lain dormant as people wrote their "What Is Computation?" position papers.

Date: Mon, 23 Nov 92 23:53:45 EST From: "Mail Delivery Subsystem"

From: Selmer Bringsjord Date: Mon, 23 Nov 92 23:01:50 -0500

Some thoughts on your thoughts (and as you read remember what a magnanimous step I have taken in agreeing to minimize the appearance of the formal in my thoughts :) ...):

>sb> We have the methodological difference to which you allude in your
>sb> previous note. And we have a difference, perhaps, over the conditional
>sb> at the heart of Turing's case for TT. Let me try now not only to
>sb> address the second, but the first also - in one stroke.
>sb> The conditional, in general, is simply the main thesis which Turing
>sb> advanced, and which many thinkers since have likewise promoted: if
>sb> something x passes the TT, then x is conscious. It was never enough
>sb> just to state the TT and pack up and go home; *Mind* wouldn't have
>sb> accepted the paper in that case. Your TTT and TTTT are intrinsically
>sb> interesting, but what excites people about them is that perhaps *they*
>sb> can supplant TT in Turing's conditional to yield a true proposition!

>
>sh> Fine, but let me immediately add a clarification. I think it was
>
>sh> arbitrary for Turing to formulate his criterion in the affirmative.
>
>sh> The correct version would be: If a candidate passes the TT we are no more
>
>sh> (or less) justified in denying that it has a mind than we are in the
>

>sh> case of real people. That’s the interesting methodological thesis
>
>sh> (false in the case of the TT, by my lights, but true in the case of the
>
>sh> TTT, again by my lights, and overdetermined in the case of the TTTT)
>
>sh> that I, at any rate, find worthy of empirical testing and logical
>
>sh> analysis. Any stronger thesis just increases the quantity of
>
>sh> arbitrariness in the problem, in my view.

Your rendition of Turing’s original conditional -- call it TT-P* -- is one that’s likely to have him turning in his grave ... because TT- P* is obviously false: Contenders for a solution to the problem of other minds involve reference to the physical appearance and behavior of real people over and above their linguistic comm -- and TT, as we all know, prohibits such reference. So we didn’t need Searle’s CRA. We can just use your construal and this shortcut. (You know, the referees at *Mind* would have picked this up. So it’s a good thing Turing didn’t propose your TT-P*. If TT-P* rests, on the other hand, on the claim that we often hold that x is a person on evidence weaker than that found in TT, you have delivered a parody of Turing’s position -- since I for one sometimes hold that x is a person on the strength of a fleeting glance at an image in a mirror, or a dot on the horizen, etc).

Of course, since I think my argument from serendipity (in Ford & Glymour, below) also refutes Turing’s original conditional (perhaps in whatever form it takes), I don’t worry much about the position you now find yourself occupying (and of course *you* won’t worry much, since you want "robotic" elements in the brew, and I seem to have thrown a few in). But you better strap your seat belt on, or at least ratchet it down tighter than it’s been during your defense of CRA, ’cause we both agree that a lot of ostensibly clever people have based their intellectual lives in large part on Turing’s original, bankrupt conditional -- and my prediction is that you’re gonna eventually find yourself coming round to extensions of CRA which shoot down that remnant of "Strong" AI that’s still near and dear to your heart.

>sb> It’s the same situation with Turingish tests which can’t in principle
>sb> be passed by finite state automata, or Turing machines: the tests
>sb> themselves are interesting, but the key is that they are supposed to
>sb> help Turing’s case. (This point is made especially vivid by the fact
>sb> that people have proposed tests which no physical artifact could pass -
>sb> on the reasonable assumption that machines beyond TMs will be, to put
>sb> it mildly, rather hard to build. Certainly such tests are devised only
>sb> to produce a defensible conditional (they may not succeed), not to give
>sb> rise to a concrete empirical goal toward which AI should strive.

>
>sh> Quite so; and I, for one, have no particular interest in defensible
>
>sh> conditionals with no empirical content or consequences -- not on
>
>sh> this topic, at any rate.

Hmm. This topic, any philosophical topic, I treat as in some sense just a branch of logic and mathematics. (Had it not been for the fact that conceding Cantor's paradise wasn't all that "expensive," we'd prolly still have thinkers around today looking for desperate escapes. The dictum that "a proof is a proof" is silly.) Your "not on this topic" suggests that had you my attitude you would take purely formal philosophical conditionals seriously. At any rate, the conditionals in question are hardly devoid of empirical content. Take "If x is a person capable of evaluating a novel like *War & Peace*, then x can decide a Turing undecidable set." This has strong empirical content for an AInik who thinks that no physical artifact can decide Turing undecidable sets! On the other hand, if you have some idiosyncratic construal waiting in the wings again (this time for 'empirical content'), what is it?

>sb> But both
>sb> of us do view TT and the like to be in part an empirical goal worth
>sb> shooting for. At any rate, I have expressed the rough-and-ready
>sb> conditional in English; it isn't formalized. In *general* such
>sb> informality, coming at a crucial dialectical juncture, worries me;
>sb> in *general* the informality is something you find welcome. But I
>sb> think we have here a case where the informal is unfortunate, as the
>sb> following reasoning may show.
>sb>
>sb> The simplest construal of Turing's conditional (sparked by Professor
>sb> Jefferson's Lister Oration in the original paper) is
>sb>
>sb> (TT-P) For every x, if x passes TT, then x is conscious.
>sb>
>sb> where the if-then here is the material conditional. But (TT-P) cannot
>sb> be what is intended, since on standard model-theoretic semantics
>sb> for first-order logic (TT-P) is vacuously true. Because (TT-P) then
>sb> says that for every element of the domain, if it passes TT, then it is
>sb> conscious - and the if-then here is again the material conditional.
>sb> Since no element of the domain passes TT, the antecedent is
>sb> always false, and therefore by the characteristic truth-table for the
>sb> material conditional the conditional is true.

>
>sh> Still not English enough. I assume what you mean is that if
>
>sh> anything OTHER THAN US passes the TT, then it's conscious -- and nothing
>
>sh> other than us passes the TT, so the claim is trivially true.

Exactly.

>
>sh> But if we
>
>sh> construe this empirically, there may eventually be something other than
>
>sh> us that passes the TT, and there are already coherent things we can say

>
>sh> even about that hypothetical future contingency (e.g., Searle's
>
>sh> Chinese Room Argument and my Symbol Grounding Problem).

Yes, but whatever you want to say about hypothetical futures and the like will suggest to me we won't really know be in position to assign truth-values until we turn to some logic for help. A large part of philosophy has been and continues to be devoted to schemes for talking carefully about hypothetical futures. This is why in an attempt to rescue Turing people turn to the many and varied conditionals of conditional logic, probabilstic conditionals, etc.

>sb> On the other hand, neither can this conditional work:
>sb>
>sb> (TT-P') It's logically necessary that (TT-P).
>sb>
>sb> The operator here I wrote before as 'L' for the necessity operator
>sb> in modal logics, more often a box. Now (TT-P') is overthrown (in
>sb> standard modal contexts, i.e., if the necessity and possibility operators
>sb> have standard meanings cashed out by normal systems of modal logic)
>sb> by what we both affirm, viz.,
>sb>
>sb> (1) It's logically possible that some thing can pass TT but not be
>sb> conscious.
>sb>
>sb> The proof that (1) implies not-(TT-P') is elementary, and as you
>sb> say, we don't want it in this context anyway.

>
>sh> Not only is it elementary, but no "proof" is necessary, because, as
>
>sh> I said earlier, the affirmative thesis is too strong, indeed it's an
>
>sh> arbitrary claim. The only thing that would have given the TT the force
>
>sh> force of necessity would have been a PROOF that anything that passed
>
>sh> it had to be conscious. No one has even given a hint of such a proof,
>
>sh> so it's obvious that the thesis is not stating a necessary truth.

Your reasoning here is fallacious, since it assumes that if P is a necessary truth, someone has given at least the hint of a proof of P. Counter-examples: Propositions like "The equivalence of deterministic and non-deterministic TMs added by disjunction introduction to the proposition that Quayle's IQ is greater than 3." Also, the set of necessary truths is uncountable.

>
>sh> In its affirmative form it is an over-strong and probably untestable
>
>sh> empirical hypothesis (which, like all empirical hypotheses, depends on

>
>sh> future data that can be adduced for and against it) and in its proper
>
>sh> negative form it is merely a methodological observation (perhaps also
>
>sh> open to counterevidence or logical counter-examples, only I haven't
>
>sh> seen any successful instances of either yet).

You have only to formalize the aforementioned objection to TT-P* via commonalities in proposed solutions to the problem of other minds.

>sb> The subjunctive approach, at least the simplest version of it, is to
>sb> augment (TT-P) with some such thing as: TT must be such that
>sb> had it
>sb> been run in other circumstances at other times, it would *also* be
>sb> passed. Here again, though, the question of counterfactuals has
>sb> occupied and continues to occupy logicians and philosophers of
>sb> science and language - and there's no consensus. And anyway I try to
>sb> show in the paper that on the dominant view of counterfactuals Turing's
>sb> subjunctivized conditional is falsified by the argument from
>sb> serendipity.

>
>sh> Look, let's make it simpler. Here's a TT (actually a TTT) for an
>
>sh> airplane: If it has performance capacity Turing-indistinguishable
>
>sh> from that of an airplane, it flies. No subjunctives needed. Of course you
>
>sh> can gerrymander the test of its flying capacity for a finite amount of
>
>sh> time to make it appear as if it can fly, even though it really can't;
>
>sh> perhaps, with sufficient control of every other physical object
>
>sh> and force for the rest of time (by God or by Chance -- "serendipity")
>
>sh> you could do it forever. So what? We're not interested in tricks, and
>
>sh> the issue is not one of necessity or of subjunctives. There's a certain
>
>sh> functional capacity a plane has, we call that flying, and anything
>
>sh> else with the same functional capacity also flies. There is no higher
>
>sh> authority.

There’s no certified logic of analogy for use in deductive contexts. Witness the problems people encounter when they try to reason to the conclusion that people are fundamentally computers because computers behave analogously to people (Chap II, WRC&CB). People are people; planes are planes. It does you no good to prove a version of TT-P* in which planes are substituted for people.

Never said I *was* interested in tricks. We’re both working toward building TT and TTT passers. But you’re forgetting what Searle has taught you. Your view mirrors the sort of desperation we find amongst those who reject the near-proof of CRA! If you tell me that a capacity for TTT-passing is what we need for consciousness, and I promptly build you a thought-experiment in which this capacity is around in all its glory, but no one’s home -- well, if you don’t get a little worried, you look like those who tell me that Jonah doing his thing with mental images of Register Machines has given birth to numerically distinct persons. I can prove, I believe, that a number of Turingish conditionals regarding TT and TTT are false (done in the relevant paper). You have simply provided a new conditional -- TT-P*; the CRAless attack on which is sketched above - - which presumably can be adapted for TTT:

(TTT-P*) If a candidate passes TTT we are no more (or less) justified in denying that it has a mind then we are in the case of real people.

I think this proposition is demonstrably false. And I’m not gonna have to leave my armchair to muster the counter-argument.

>sb> Everything I’ve said is of course consistent with my position that
>sb> our robots will pass TT, TTT, ..., but will not be persons. In the
>sb> present context, you could say that this is partly because persons are
>sb> essentially (self-) conscious, and standardly conceived robots
>sb> aren’t. (Chapter IX in WRC&CB is an argument for this.)

>
>sh> I would not say anything of the sort.

I knew that. (But what premise is false in Chapter IX?)

>
>sh> I think it is possible that all TTT-passers will be conscious and I
>
>sh> think it is possible that not all TTT-passers will be conscious. I just
>
>sh> think there are methodological reasons, peculiar to the mind/body
>
>sh> problem, why we can never know one way or the other, even with the
>
>sh> normal level of uncertainty of an ordinary empirical hypothesis. (By
>
>sh> the way, although I believe that Searle’s Chinese Room Argument has
>
>sh> shown that it is highly improbable that TT-passing computers are
>

>sh> conscious, it is not a PROOF that they cannot be; a second
>
>sh> consciousness of which Searle is not conscious is still a logical
>
>sh> possibility, but not one worthy of much credence, in my view.)

Come on, Stevan, I'm tellin' you, check those seat belts. My Searlean argument in *What Robots Can & Can't Be* is pretty close to a proof. The David Cole/ (I now see:) Michael Dyer multiple personality desperation dodge is considered therein, the upshot being that LISP programmers have the capacity to throw off the Earth's census. I've tried this out on students for five years. If you take my chapter and teach it and take polls before and after, the results don't bode well for "Strong" AI. You know you think the multiple person move implies some massively counter-intuitive things'll need to be swallowed. And the Hayes objection is easily handled by my Jonah, who works at the level of Register Machines and rocks -- besides which, Jonah can be hypnotized etc. etc. so as to remove "free will" from the picture (a move I spelled out for Hayes' after his presentation of his CRA dodge at the Second Human & Machine Cognition workshop). Besides, again, you follow herdish naivete and talk as if a proof is a proof -- in some religious sense. My version of CRA is as much of a proof as any reductio in classical mathematics is for a constructivist. Intuitionist mathematics is perfectly consistent, clever, rigorous, has been affirmed by some biggies, and so on. Try getting an intuitionist to affirm some of the non-constructive theorems which most in this forum seem to be presupposing. I can prove Goldbach's conjecture is either T or F in a sec in Logic 101 But that's not a proof for a genius like Heyting! You want to rest on probability (w.r.t. CRA etc.). I want proofs. But the motivation to rest on proofs may come from a mistaken notion of 'proof.'

>sb> Right now we're covering the symbol grounding problem in my
>sb> course Philosophy of AI (upper undergraudate level), and my students
>sb> greatly enjoy your writings on the matter. We may arrive at something
>sb> worth sending on to you. Were looking now at whether some such
>sb> proposal as "a formula (P) in some robot R's KB means P for R iff
>sb> some causal relation obtains between the R's sensors and effectors, the
>sb> external physical world, and KB" is promising. This is a proposal which
>sb> Georges Rey has made, and it seems related to your proposal for how to
>sb> ground symbols.

>
>sh> As you know, I haven't much of a taste for such formalism. If what
>
>sh> you/he are saying is that it is more probable that a symbol system is
>
>sh> conscious if it can not only pass the TT, but also the TTT, that is
>
>sh> indeed what I was saying too.

What I'm saying is that your position on TTT can be put in the form of a quasi-formal declarative proposition about the desired causal relation between symbol systems, sensors and effectors, and the physical environment. (If it *can't* be so put, I haven't much of a taste for it.) Such a proposition is sketched by Rey. Again, I'm not gonna have to leave my leather Chesterfield to come up with a diagnosis.

By all means, post our exchange to this point, and subsequent discussion if you like. Keep in mind, however, that I'm gonna have to write my position paper after finishing the archives! (So I may not be discussing much for a while.) The only reason I might be able to turn this thing around fast is that in '88 I formulated a Chalmersian definition of 'x is a computer' after going without sleep for two days, obsessed with the Searlean "everything is a computer" argument that at that time wasn't associated with Searle (it was considered by R. J. Nelson).

(Has anybody else read the archives from start to finish in hard copy having not seen any of the discussion before? You've done a marvelous job with this, Stevan. I must say, though, that after my sec'y got the files and generated hard copy within the confines of one day, she was weary: how many trees have you guys killed?

All the best, Stevan, and thanks for the stimulating discussion. I hope to have the THINK reaction finished before Turkey day, will send to you direct...

Yours, Selmer

REFERENCES

Bringsjord (in press) "Could, How Could We Tell If, and Why Should -- Androids Have Inner LIves," in Ford, K. & C. Glymour, eds., *Android Epistemology* (Greenwich, CT: JAI Press).

Bringsjord, S. (1992) *What Robots Can & Can't Be* (Dordrecht, The Netherlands: Kluwer). "Searle: Chapter V" "Introspection: Chapter IX"

From harnad Mon Jun 14 22:43:52 1993 To: brings@rpi.edu (Selmer Bringsjord) Subject: Long-overdue reply to Bringsjord Status: RO

As this response to Selmer Brinsgjord (sb) is long (over 7 months) overdue, I (sh) back-quote in extenso to furnish the context:

> From: Selmer Bringsjord > Date: Mon, 23 Nov 92 23:01:50 -0500

>sb> Some thoughts on your thoughts (and as you read remember what a
>sb> magnanimous step I have taken in agreeing to minimize the
>sb> appearance of the formal in my thoughts :) ...):

>
>sb> We have the methodological difference to which you allude in your
>
>sb> previous note. And we have a difference, perhaps, over the conditional
>
>sb> at the heart of Turing's case for TT. Let me try now not only to
>
>sb> address the second, but the first also - in one stroke.
>
>sb> The conditional, in general, is simply the main thesis which Turing
>

>sb> advanced, and which many thinkers since have likewise promoted: if
>
>sb> something x passes the TT, then x is conscious. It was never enough
>
>sb> just to state the TT and pack up and go home; *Mind* wouldn't have
>
>sb> accepted the paper in that case. Your TTT and TTTT are intrinsically
>
>sb> interesting, but what excites people about them is that perhaps *they*
>
>sb> can supplant TT in Turing's conditional to yield a true proposition!

>
>sh> Fine, but let me immediately add a clarification. I think it was
>
>sh> arbitrary for Turing to formulate his criterion in the affirmative.
>
>sh> The correct version would be: If a candidate passes the TT we are no more
>
>sh> (or less) justified in denying that it has a mind than we are in the
>
>sh> case of real people. That's the interesting methodological thesis
>
>sh> (false in the case of the TT, by my lights, but true in the case of the
>
>sh> TTT, again by my lights, and overdetermined in the case of the TTTT)
>
>sh> that I, at any rate, find worthy of empirical testing and logical
>
>sh> analysis. Any stronger thesis just increases the quantity of
>
>sh> arbitrariness in the problem, in my view.

>sb> Your rendition of Turing's original conditional -- call it TT-P* -- is
>sb> one that's likely to have him turning in his grave ... because TT- P*
>sb> is obviously false: Contenders for a solution to the problem of other
>sb> minds involve reference to the physical appearance and behavior of real
>sb> people over and above their linguistic comm -- and TT, as we all know,
>sb> prohibits such reference. So we didn't need Searle's CRA. We can just
>sb> use your construal and this shortcut. (You know, the referees at *Mind*
>sb> would have picked this up. So it's a good thing Turing didn't propose
>sb> your TT-P*.

No, I think I have Turing's actual intuition (or what ought to have been his intuition) right, and it's not obviously false -- despite appearances, so to speak: Turing of course knew we use appearances, but he also knew appearances can be deceiving (as they would be if a computer really DID have a mind but we immediately dismissed it out of hand simply because of the way it looked). The point of Turing's party game was partly to eliminate BIAS based on appearance. We would certainly be prepared to believe that other organisms, including extraterrestrial ones, might have minds, and we have no basis for legislating in advance what their exteriors are or are not allowed to LOOK like (any more than we can legislate in advance what their interiors are supposed to look like). It’s what they can DO that guides our judgment, and it’s the same with you and me (and the Blind Watchmaker: He can’t read minds either, only adaptive performance).

I have no idea how your brain works, and my intuitions about appearances are obviously anthropocentric and negotiable. But if you could correspond with me as a lifelong pen-pal I have never seen, in such a way that it would never cross my mind that you had no mind, then it would be entirely arbitrary of me to revise my judgment just because I was told you were a machine -- for the simple reason that I know neither what PEOPLE are nor what MACHINES are. I only know what machines and people usually LOOK like, and those appearances can certainly be deceiving. By contrast, the wherewithal to communicate with me intelligibly for a lifetime -- that’s something I understand, because I know exactly where it’s coming from, so to speak.

On the other hand, I of course agree that what a person can DO includes a lot more than pen-pal correspondence, so in eliminating bias based on appearance, Turing inadvertently and arbitrarily circumscribed the portion of our total performance capacity that was to be tested. It is here that my own upgrade from T2 to T3 is relevant. It should be noted, though, that T3 is still in the spirit of Turing’s original intuition. It’s not really the APPEARANCE of the candidate that is relevant there either, it is only what it can DO -- which includes interacting with the objects in the world as we do (and, of course, things like eye contact and facial expression are probably important components of what we can and do DO too, but that’s another matter).

So, yes, restriction to T2 rather than T3 was a lapse on Turing’s part, though this still did not make even T2 FALSE prima facie: We still needed a nonarbitrary REASON for revising our intuitions about the fact that our lifelong pen-pal had a mind even if he turned out to be a machine, and this revision could not be based on mere APPEARANCE, which is no reason at all. So Searle’s Chinese Room Argument WAS needed after all, to show -- at least in the case of a life-long pen-pal who was ONLY the implementation of an implementation-independent symbol system, each of whose implementations allegedly had a mind -- that the conclusion drawn from T2 would have been false.

But note that I say "would have been," since I do not believe for a minute that a pure symbol-cruncher could ever actually pass T2 (because of the symbol grounding problem). I believe a successful T2 candidate’s capacity to pass T2 would have to be grounded in its capacity to pass T3 -- in other words, the candidate would still have to be a robot, not just a symbol-cruncher. So, paradoxically, T2 is even closer to validity than might appear even once we take into account that it was arbitrarily truncated, leaving out all the robotic capacities of T3; for if I’m right, then T2 is also an indirect test of T3. Only for a counterfactual symbol-cruncher that could hypothetically pass T2 is T2 not valid (and so we do need Searle’s Argument after all).

To summarize: the only nonarbitrary basis we (or the Blind Watchmaker) have for adjudicating the presence or absence of mind is performance capacity totally indistinguishable from that of entities that really do have minds. External appearances -- or internal ones, based on the means by which the performance capacity is attained -- are irrelevant because we simply KNOW nothing about either.

And I repeat, no tricks are involved here. We REALLY want the performance capacity; that's not just an intuitive criterion but an empirical, engineering constraint, narrowing the degrees of freedom for eligible candidates to the ones that will (we hope) admit only those with minds. For this reason it is NOT a valid objection to point out that, for example, people who are paralyzed, deaf and blind still have minds. Of course they do; but the right way to converge on the correct model is not to look first for a candidate that is T3-indistinguishable from such handicapped persons, but for one that is T3-indistinguishable from a normal, intact person. It's after that's accomplished that we can start worrying about how much we can pare it back and still have somebody home in there.

A more relevant objection was my hint at the adaptive role of appearance in facial expression, for example, but my guess is that this will be a matter fine-tuning for the winning candidate (as will neurosimilitude). The empirical degrees of freedom are probably narrowed sufficiently by our broad-stroke T3 capacity.

>sb> If TT-P* rests, on the other hand, on the claim that we
>sb> often hold that x is a person on evidence weaker than that found in TT,
>sb> you have delivered a parody of Turing's position -- since I for one
>sb> sometimes hold that x is a person on the strength of a fleeting glance
>sb> at an image in a mirror, or a dot on the horizen, etc).

No, as I have written (Harnad 1992), the right construal of T2 is as a life-long pen-pal. Short-term party tricks and snap conclusions are not at issue here. We're talking about an empirical, engineering criterion. By way of analogy, if the goal is to build a system with performance capacities indistinguishable from those of an airplane, it is not good enough to build something that will fool you into think it's a plane for a few minutes, or even hours. It REALLY has to have a plane's total performance capacity.

>sb> Of course, since I think my argument from serendipity (in Ford &
>sb> Glymour, below) also refutes Turing's original conditional (perhaps in
>sb> whatever form it takes), I don't worry much about the position you
>sb> now find yourself occupying (and of course *you* won't worry much,
>sb> since you want "robotic" elements in the brew, and I seem to have
>sb> thrown a few in).

Regarding the serendipity argument, let me quote from my reply to your commentary (Bringsjord 1993) on Harnad (1993a): "So chimpanzees might write Shakespeare by chance: What light does that cast on how Shakespeare wrote Shakespeare?" This is perhaps the difference between an engineering and a philosophical motivation on this topic. Physicists don't worry about thermodynamics reversing by chance either, even though it's a logical possibility. It's just that nothing substantive rides on that possibility. A life-long pen-pal correspondence could be anticipated by chance, logically speaking. So what? What we are looking for is a mechanism that does it by design, not by chance.

My own position is that T3 is the right level of empirical constraint for capturing the mind (though it is of course no guarantor). T2 is too UNDERdetermined (because it lets in the hypothetical ungrounded symbol-cruncher) and T4 is OVERconstrained (because we do not know which of our neuromolecular properties are RELEVANT to generating our T3 capacity). And that's all there is to the reverse-engineering T-hierarchy ("t1" is subtotal "toy" fragments of our Total capacity, hence not a Turing Test at all, and hopelessly underdetermined, and T5 is the Grand Unified Theory of

Everything, of which engineering and reverse engineering -- T2 - T4 -- are merely a fragment; Harnad 1994). There's just nowhere else to turn empirically; and no guarantees exist. That's what makes the mind/body problem special, with an extra layer of underdetermination, over and above that of pure physics and engineering: Even in T4 there could be nobody home, for all we know. This makes qualia fundamentally unlike, say, quarks, despite the fact that both are undeservable and both are real (Harnad 1993b).

>sb> But you better strap your seat belt on, or at least
>sb> ratchet it down tighter than it's been during your defense of CRA,
>sb> 'cause we both agree that a lot of ostensibly clever people have
>sb> based their intellectual lives in large part on Turing's original,
>sb> bankrupt conditional -- and my prediction is that you're gonna
>sb> eventually find yourself coming round to extensions of CRA which
>sb> shoot down that remnant of "Strong" AI that's still near and dear to
>sb> your heart.

I'm all strapped in, but the view from here is that the only thing that made CRA work, the only thing that gave Searle this one special periscope for peeking across the otherwise impenetrable other-minds barrier, is the implementation-independence of pure symbol systems. Hence it is only a T2-passing symbol cruncher that is vulnerable to CRA. There are no extensions. T3 is as impenetrable as a stone or a brain.

>
>sb> It's the same situation with Turingish tests which can't in principle
>
>sb> be passed by finite state automata, or Turing machines: the tests
>
>sb> themselves are interesting, but the key is that they are supposed to
>
>sb> help Turing's case. (This point is made especially vivid by the fact
>
>sb> that people have proposed tests which no physical artifact could pass -
>
>sb> on the reasonable assumption that machines beyond TMs will be, to put
>
>sb> it mildly, rather hard to build. Certainly such tests are devised only
>
>sb> to produce a defensible conditional (they may not succeed), not to give
>
>sb> rise to a concrete empirical goal toward which AI should strive.

>
>sh> Quite so; and I, for one, have no particular interest in defensible
>
>sh> conditionals with no empirical content or consequences -- not on
>
>sh> this topic, at any rate.

>sb> Hmm. This topic, any philosophical topic, I treat as in some sense just
>sb> a branch of logic and mathematics. (Had it not been for the fact that
>sb> conceding Cantor’s paradise wasn’t all that "expensive," we’d prolly
>sb> still have thinkers around today looking for desperate escapes.
>sb> The dictum that "a proof is a proof" is silly.) Your "not on this
>sb> topic" suggests that had you my attitude you would take purely formal
>sb> philosophical conditionals seriously. At any rate, the conditionals in
>sb> question are hardly devoid of empirical content. Take "If x is a person
>sb> capable of evaluating a novel like *War & Peace*, then x can decide a
>sb> Turing undecidable set." This has strong empirical content for an AInik
>sb> who thinks that no physical artifact can decide Turing undecidable
>sb> sets! On the other hand, if you have some idiosyncratic construal
>sb> waiting in the wings again (this time for ’empirical content’), what is
>sb> it?

Nothing idiosyncratic. We are interested in real reverse engineering here: Designing systems that really have certain capacities. If there are arguments (like CRA) that show that certain approaches to this reverse engineering (like trying to design a T2-scale symbol-cruncher) are likely to fail, then they are empirically relevant. Otherwise not. Engineering is not a branch of logic and mathematics (though it may apply them, and is certainly bound by them). Conditionals about undecidability are (as far as I can see) irrelevant; we are interested in capturing what people CAN do (T3), not what they can’t...

>
>sb> But both
>
>sb> of us do view TT and the like to be in part an empirical goal worth
>
>sb> shooting for. At any rate, I have expressed the rough-and-ready
>
>sb> conditional in English; it isn’t formalized. In *general* such
>
>sb> informality, coming at a crucial dialectical juncture, worries me;
>
>sb> in *general* the informality is something you find welcome. But I
>
>sb> think we have here a case where the informal is unfortunate, as the
>
>sb> following reasoning may show.

>
>sb> The simplest construal of Turing’s conditional (sparked by Professor
>
>sb> Jefferson’s Lister Oration in the original paper) is
>
>sb> (TT-P) For every x, if x passes TT, then x is conscious.
>
>sb> where the if-then here is the material conditional. But (TT-P) cannot
>

>sb> be what is intended, since on standard model-theoretic semantics
>
>sb> for first-order logic (TT-P) is vacuously true. Because (TT-P) then
>
>sb> says that for every element of the domain, if it passes TT, then it is
>
>sb> conscious - and the if-then here is again the material conditional.
>
>sb> Since no element of the domain passes TT, the antecedent is
>
>sb> always false, and therefore by the characteristic truth-table for the
>
>sb> material conditional the conditional is true.

>
>sh> Still not English enough. I assume what you mean is that if
>
>sh> anything OTHER THAN US passes the TT, then it's conscious -- and nothing
>
>sh> other than us passes the TT, so the claim is trivially true.

>sb> Exactly.

>
>sh> But if we
>
>sh> construe this empirically, there may eventually be something other than
>
>sh> us that passes the TT, and there are already coherent things we can say
>
>sh> even about that hypothetical future contingency (e.g., Searle's
>
>sh> Chinese Room Argument and my Symbol Grounding Problem).

>sb> Yes, but whatever you want to say about hypothetical futures and the
>sb> like will suggest to me we won't really be in position to assign
>sb> truth-values until we turn to some logic for help. A large part of
>sb> philosophy has been and continues to be devoted to schemes for talking
>sb> carefully about hypothetical futures. This is why in an attempt to
>sb> rescue Turing people turn to the many and varied conditionals of
>sb> conditional logic, probabilistic conditionals, etc.

The only conditional I've found useful here so far is Searle's: "IF a pure symbol-cruncher could pass T2, it would not understand, because I could implement the same symbol-cruncher without understanding."

>
>sb> On the other hand, neither can this conditional work:

>
>sb> (TT-P') It's logically necessary that (TT-P).
>
>sb> The operator here I wrote before as 'L' for the necessity operator
>
>sb> in modal logics, more often a box. Now (TT-P') is overthrown (in
>
>sb> standard modal contexts i.e. if the necessity and possibility operators
>
>sb> have standard meanings cashed out by normal systems of modal logic)
>
>sb> by what we both affirm, viz.,
>
>sb> (1) It's logically possible that some thing can pass TT but not be
>
>sb> conscious.
>
>sb> The proof that (1) implies not-(TT-P') is elementary, and as you
>
>sb> say, we don't want it in this context anyway.

>
>sh> Not only is it elementary, but no "proof" is necessary, because, as
>
>sh> I said earlier, the affirmative thesis is too strong, indeed it's an
>
>sh> arbitrary claim. The only thing that would have given the TT the
>
>sh> force of necessity would have been a PROOF that anything that passed
>
>sh> it had to be conscious. No one has even given a hint of such a proof,
>
>sh> so it's obvious that the thesis is not stating a necessary truth.

>sb> Your reasoning here is fallacious, since it assumes that if P is a
>sb> necessary truth, someone has given at least the hint of a proof of P.
>sb> Counter-examples: Propositions like "The equivalence of deterministic
>sb> and non-deterministic TMs added by disjunction introduction to the
>sb> proposition that Quayle's IQ is greater than 3." Also, the set of
>sb> necessary truths is uncountable.

I can only repeat, "For every x, if x passes TT, then x is conscious" was just a conjecture, with some supporting arguments. It turns out to be very probably false (unless memorizing symbols can make Searle understand Chinese, or generate a second mind in him that understands Chinese). No Searlean argument can be made against T3, but that could of course be false too; and so could even T4. So forget about proofs or necessity here. It's underdetermination squared all the way through, because of what's abidingly special about the mind/body problem (or about qualia, if you prefer).

>
>sh> In its affirmative form it is an over-strong and probably untestable
>
>sh> empirical hypothesis (which, like all empirical hypotheses, depends on
>
>sh> future data that can be adduced for and against it) and in its proper
>
>sh> negative form it is merely a methodological observation (perhaps also
>
>sh> open to counterevidence or logical counter-examples, only I haven't
>
>sh> seen any successful instances of either yet).

>sb> You have only to formalize the aforementioned objection to TT-P* via
>sb> commonalities in proposed solutions to the problem of other minds.

I cannot understand it stated in that abstract formal form. In plain English, how does it refute that T3 is the right level of empirical constraint for capturing the mind (bearing in mind that T3 never guaranteed anything, and never could, since nothing could) in the reverse engineering sense?

>
>sb> The subjunctive approach, at least the simplest version of it, is to
>
>sb> augment (TT-P) with some such thing as: TT must be such that
>
>sb> had it
>
>sb> been run in other circumstances at other times, it would *also* be
>
>sb> passed. Here again, though, the question of counterfactuals has
>
>sb> occupied and continues to occupy logicians and philosophers of
>
>sb> science and language - and there's no consensus. And anyway I try to
>
>sb> show in the paper that on the dominant view of counterfactuals Turing's
>
>sb> subjunctivized conditional is falsified by the argument from
>
>sb> serendipity.

>
>sh> Look, let's make it simpler. Here's a TT (actually a TTT) for an
>
>sh> airplane: If it has performance capacity Turing-indistinguishable
>
>sh> from that of an airplane, it flies. No subjunctives needed. Of course
>
>sh> you can gerrymander the test of its flying capacity for a finite amount

>
>sh> of time to make it appear as if it can fly, even though it really can't;
>
>sh> perhaps, with sufficient control of every other physical object
>
>sh> and force for the rest of time (by God or by Chance -- "serendipity")
>
>sh> you could do it forever. So what? We're not interested in tricks, and
>
>sh> the issue is not one of necessity or of subjunctives. There's a certain
>
>sh> functional capacity a plane has, we call that flying, and anything
>
>sh> else with the same functional capacity also flies. There is no higher
>
>sh> authority.

>sb> There's no certified logic of analogy for use in deductive contexts.
>sb> Witness the problems people encounter when they try to reason to the
>sb> conclusion that people are fundamentally computers because computers
>sb> behave analogously to people (Chap II, WRC&CB). People are people;
>sb> planes are planes. It does you no good to prove a version of TT-P* in
>sb> which planes are substituted for people.

I can't follow this. Forward engineering was able to build and completely explain the causal principles of planes. I'm just suggesting that reverse engineering T3 capacity will do the same with people, and, en passant, it will also capture the mind. The point is not about analogy, it's about the generation of real performance capacities.

>sb> Never said I *was* interested in tricks. We're both working toward
>sb> building TT and TTT passers. But you're forgetting what Searle has
>sb> taught you. Your view mirrors the sort of desperation we find amongst
>sb> those who reject the near-proof of CRA! If you tell me that a capacity
>sb> for TTT-passing is what we need for consciousness, and I promptly build
>sb> you a thought-experiment in which this capacity is around in all its
>sb> glory, but no one's home -- well, if you don't get a little worried,
>sb> you look like those who tell me that Jonah doing his thing with mental
>sb> images of Register Machines has given birth to numerically distinct
>sb> persons. I can prove, I believe, that a number of Turingish
>sb> conditionals regarding TT and TTT are false (done in the relevant
>sb> paper). You have simply provided a new conditional -- TT-P*; the
>sb> CRAless attack on which is sketched above - - which presumably can be
>sb> adapted for TTT:
>sb> (TTT-P*) If a candidate passes TTT we are no more (or less) justified
>sb> in denying that it has a mind then we are in the case of real people.
>sb> I think this proposition is demonstrably false. And I'm not gonna have
>sb> to leave my armchair to muster the counter-argument.

I'm ready for the counter-argument. I know what made the CRA work (penetrability of the other-minds barrier to Searle's periscope because the implementation-independence of the symbolic level), but since a T3-passer cannot be just a symbol-cruncher, how is the T3-CRA argument to go through? Serendipity and Jonah are just arguments for the logical POSSIBILITY that a T2 or T3 passer should fail to have a mind. But I've never claimed otherwise, because I never endorsed the positive version of T2 or T3! For me, they are epistemic, not ontic criteria (empirical constraints on models, actually).

>
>sb> Everything I've said is of course consistent with my position that
>
>sb> our robots will pass TT, TTT, ..., but will not be persons. In the
>
>sb> present context, you could say that this is partly because persons are
>
>sb> essentially (self-) conscious, and standardly conceived robots
>
>sb> aren't. (Chapter IX in WRC&CB is an argument for this.)

>
>sh> I would not say anything of the sort.

>sb> I knew that. (But what premise is false in Chapter IX?)

I mean I never say much of anything about "self-consciousness": consciousness simpliciter (somebody home) is enough for me. Nor do I know what "standardly conceived robots" are, but the only ones I've ever had in mind are T3 robots; and although it's POSSIBLE that they will not be conscious, there is (unlike in the case of T2, symbol-crunching, and Searle's periscope) no way of being any the wiser about that, one way or the other. Repeating logical possibilities in ever more embellished parables does not advance us in either direction: The logical possibility is there; the means of being any the wiser is not.

So we may as well stick with T3, which was good enough for the original designer. The only open question is whether there is anything more to be said for T4 (the last possible empirical resort). Until further notice, I take T4 to be just a matter of fine tuning; passing T3 will already have solved all the substantive problems, and if the fact of the matter is that that's not a fine enough filter to catch the mind, whereas T4 is, then what's peculiar to mind is even more peculiar than it had already seemed: Certainly no one will ever be able to say not only WHETHER, but even, if so, WHY, among several T3-indistinguishable candidates, only the T4-indistinguishable one should have a mind.

Chapter IX is too much of a thicket. The CRA is perspicuous; you can say what's right or wrong with it in English, without having to resort to formalisms or contrived and far-fetched sci-fi scenarios. I have restated the point and why and how it's valid repeatedly in a few words; I really think you ought to do the same.

(Also, speaking as [as far as I know] the first formulator of the T-hierarchy, there is, despite your hopeful dots after T2, T3..., only one more T, and that's T4! The Turing hierarchy (for mind-modelling purposes) ENDS there, whereas the validity of CRA begins and ends at T2.)

>
>sh> I think it is possible that all TTT-passers will be conscious and I
>
>sh> think it is possible that not all TTT-passers will be conscious. I just
>
>sh> think there are methodological reasons, peculiar to the mind/body
>
>sh> problem, why we can never know one way or the other, even with the
>
>sh> normal level of uncertainty of an ordinary empirical hypothesis. (By
>
>sh> the way, although I believe that Searle's Chinese Room Argument has
>
>sh> shown that it is highly improbable that TT-passing computers are
>
>sh> conscious, it is not a PROOF that they cannot be; a second
>
>sh> consciousness of which Searle is not conscious is still a logical
>
>sh> possibility, but not one worthy of much credence, in my view.)

>sb> Come on, Stevan, I'm tellin' you, check those seat belts. My Searlean
>sb> argument in *What Robots Can & Can't Be* is pretty close to a proof.
>sb> The David Cole/ (I now see:) Michael Dyer multiple personality
>sb> desperation dodge is considered therein, the upshot being that LISP
>sb> programmers have the capacity to throw off the Earth's census. I've
>sb> tried this out on students for five years. If you take my chapter and
>sb> teach it and take polls before and after, the results don't bode well
>sb> for "Strong" AI. You know you think the multiple person move implies
>sb> some massively counter-intuitive things'll need to be swallowed. And
>sb> the Hayes objection is easily handled by my Jonah, who works at the
>sb> level of Register Machines and rocks -- besides which, Jonah can be
>sb> hypnotized etc. etc. so as to remove "free will" from the picture (a
>sb> move I spelled out for Hayes' after his presentation of his CRA dodge
>sb> at the Second Human & Machine Cognition workshop). Besides, again, you
>sb> follow herdish naivete and talk as if a proof is a proof -- in some
>sb> religious sense. My version of CRA is as much of a proof as any
>sb> reductio in classical mathematics is for a constructivist. Intuitionist
>sb> mathematics is perfectly consistent, clever, rigorous, has been
>sb> affirmed by some biggies, and so on. Try getting an intuitionist to
>sb> affirm some of the non-constructive theorems which most in this forum
>sb> seem to be presupposing. I can prove Goldbach's conjecture is either T
>sb> or F in a sec in Logic 101 But that's not a proof for a genius like
>sb> Heyting! You want to rest on probability (w.r.t. CRA etc.). I want
>sb> proofs. But the motivation to rest on proofs may come from a mistaken
>sb> notion of 'proof.'

This has nothing to do with nonconstructive vs. constructive proof. As I understand it, CRA is not and cannot be a proof. If you can upgrade it to one, say how, in a few transparent words. Students can be persuaded in many ways; that's irrelevant. I've put my construal briefly and transparently; you should do the same.

>
\>sb> Right now we're covering the symbol grounding problem in my
>
\>sb> course Philosophy of AI (upper undergraudate level), and my students
>
\>sb> greatly enjoy your writings on the matter. We may arrive at something
>
\>sb> worth sending on to you. Were looking now at whether some such
>
\>sb> proposal as "a formula (P) in some robot R's KB means P for R iff
>
\>sb> some causal relation obtains between the R's sensors and effectors, the
>
\>sb> external physical world, and KB" is promising. This is a proposal which
>
\>sb> Georges Rey has made, and it seems related to your proposal for how to
>
\>sb> ground symbols.

>
\>sh> As you know, I haven't much of a taste for such formalism. If what
>
\>sh> you/he are saying is that it is more probable that a symbol system is
>
\>sh> conscious if it can not only pass the TT, but also the TTT, that is
>
\>sh> indeed what I was saying too.

\>sb> What I'm saying is that your position on TTT can be put in the form of
\>sb> a quasi-formal declarative proposition about the desired causal
\>sb> relation between symbol systems, sensors and effectors, and the
\>sb> physical environment. (If it *can't* be so put, I haven't much of a
\>sb> taste for it.) Such a proposition is sketched by Rey. Again, I'm not
\>sb> gonna have to leave my leather Chesterfield to come up with a
\>sb> diagnosis.

But there's no NEED for such a formalization, any more than there is for the engineering criterion of building a plane with flight capacities indistinguishable from those of a DC-11. A robot's internal symbols (if any) are grounded if it can pass T3: The interpretations of the symbols then do not need to be mediated by an interpreter; they are grounded in the robot's T3 interactions with the objects, properties, events and states of affairs in the world that the symbols are otherwise merely externally interpretable as being about.

>sb> All the best, Stevan, and thanks for the stimulating discussion.
>sb> Yours, Selmer

>sb> REFERENCES

>sb> Bringsjord (in press) "Could, How Could We Tell If, and Why Should --
>sb> Androids Have Inner LIves," in Ford, K. & C. Glymour,
>sb> eds., *Android Epistemology* (Greenwich, CT: JAI Press).

>sb> Bringsjord, S. (1992) *What Robots Can & Can't Be* (Dordrecht, The
>sb> Netherlands: Kluwer). "Searle: Chapter V" "Introspection: Chapter IX"

Thanks Selmer, and sorry for the long delay in my response! -- Stevan

----

Bringsjord, S. (1993) People Are Infinitary Symbol Systems: No Sensorimotor Capacity Necessary. Commentary on Harnad (1993) Think (Special Issue on Machine Learning) (in press)

Harnad, S. (1992) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. SIGART Bulletin 3(4) (October) 9 - 10.

Harnad, S. (1993a) Grounding Symbols in the Analog World with Neural Nets. Think (Special Issue on Machine Learning) (in press)

Harnad, S. (1993b) Discussion. In: T. Nagel (ed.) Experimental and Theoretical Studies of Consciousness. Ciba Foundation Symposium 174.

Harnad, S, (1994) Does the Mind Piggy-Back on Robotic and Symbolic Capacity? To appear in: H. Morowitz (ed.) "The Mind, the Brain, and Complex Adaptive Systems.

The above articles are retrievable by anonymous ftp from host: princeton.edu directory: pub/harnad/Harnad

----

From harnad Fri Jun 11 21:21:34 1993 To: sharnad@life.jsc.nasa.gov Subject: Preprint by S. Yee available Status: RO

> Date: Fri, 11 Jun 93 14:04:31 -0400
> From: yee@envy.cs.umass.edu
> To: harnad@princeton.edu
>
> A revised version of my paper on Turing machines, the Chinese
> room, and Godel has been accepted for publication in the philosophy journal
> LYCEUM. I was wondering whether (a) I could send the following announcement
> to the Symbol Grounding list, and (b) I could get a listing of SG-list
> subscribers so as to avoid sending any of them a duplicate announcement?
>
> Thank you very much, Richard

> Subscribers to this list might be interested in the following
> article, which will appear in LYCEUM, 5(1), Spring 1993.
>
> PostScript versions are available via anonymous ftp to
> envy.cs.umass.edu, file: pub/yee/tm-semantic.ps.Z. Instructions are
> provided below. Hard-copies are also available from the author.
>
> Questions and reactions to the article are welcomed. RY
>
>
> TURING MACHINES AND SEMANTIC SYMBOL PROCESSING:
>
> Why Real Computers Don't Mind Chinese Emperors
>

> Richard Yee
>
> Department of Computer Science
> University of Massachusetts, Amherst, MA 01003
> Internet: yee@cs.umass.edu
> Tel: (413) 545-1596, 549-1074
>

> Abstract
>

> Philosophical questions about minds and computation need to focus
> squarely on the mathematical theory of Turing machines (TM's).
> Surrogate TM's such as computers or formal systems lack abilities
> that make Turing machines promising candidates for possessors of
> minds. Computers are only universal Turing machines (UTM's)---a
> conspicuous but unrepresentative subclass of TM. Formal systems are
> only static TM's, which do not receive inputs from external sources.
> The theory of TM computation clearly exposes the failings of two
> prominent critiques, Searle's Chinese room (1980) and arguments from
> \Godel's Incompleteness theorems (e.g., Lucas, 1961; Penrose, 1989),
> both of which fall short of addressing the complete TM model. Both
> UTM-computers and formal systems provide an unsound basis for
> debate. In particular, their special natures easily foster the
> misconception that computation entails intrinsically meaningless
> symbol manipulation. This common view is incorrect with respect to
> full-fledged TM's, which can process inputs non-formally, i.e., in a
> subjective and dynamically evolving fashion. To avoid a distorted
> understanding of the theory of computation, philosophical judgements
> and discussions should be grounded firmly upon the complete Turing
> machine model, the proper model for real computers.

>
> ==
>

```
> Instructions for anonymous, binary ftp:
> ---------------------------------------

>
> unix> ftp envy.cs.umass.edu
>
> Name: anonymous
> Password:
> ftp> cd pub/yee
> ftp> binary
> ftp> get tm-semantic.ps.Z
> ftp> bye
>
> unix> uncompress tm-semantic.ps.Z
> unix> tm-semantic.ps
```

Date: Fri, 11 Jun 93 22:25:30 EDT From: "Stevan Harnad" To: yee@envy.cs.umass.edu Subject: Prima facie questions

Comments on Yee's Paper:

I could only print out page 16 and upward of Yee's paper, but from that I was able to discern the following: Once one sets aside the abstractions and technical language on the one hand, and the mentalistic interpretation on the other, Yee seems to be saying that only UTMs (Universal Turing Machines, e.g. digital computers) are symbol-manipulators, and hence open to the objections against pure symbol manipulation. TMs (Turing Machines) are not. UTMs merely SIMULATE TMs, which are in turn OTHER kinds of machines that are NOT just symbol manipulators. Systems with minds are TMs, not UTMs.

That all sounds fine, but one still needs the answers to a few questions:

(1) What kind of system is NOT a TM then, in this sense? Is a bridge, a furnace, a plane, a protein, an organism, an atom, a solar system? a brain? the universe?

(2) If the answer to all of the above is that they are all TMs, then OF COURSE systems with minds are TMs too, but then what does this tell us about the mind that isn't true about everything else under the sun (including the sun) as well? Saying it was a TM was supposed to tell us something more than that it was a physical system! We ALL accepted that in the first place. The question was, what KIND of physical system. If "TMs" does not pick out a subset, it's not informative, just as it would not be informative to tell us that the brain, like everything else, obeys differential equations.

(3) If the answer is instead that NOT all physical systems are TMs, then what distinguishes those that are from those that aren't (and what's special about the kinds with minds)? In the version of the cognition-is-computation hypothesis that I know, the UTM version, not the TM version, mentation is a form of symbol manipulation and mental states are symbolic states, independent, like software, of the physical details of their implementation. Is there implementation-independence for TMs too (it seems to me you NEED this so someone cannot say it's just differential equations, i.e., physics)? If so, WHAT is independent of the implementation, if it is not the symbol system, as in the case of UTMs? Is there multiple realizability too (i.e., would EVERY implementation of whatever it is that is

implementation-independent in a TM [it's not symbols and syntax any more, so what is it?] have a mind if we found the right TM?)? Would the UTM simulating that TM have a mind? And if not, why not?

These are the kinds of things you have to be explicit about, otherwise your reader is lost in an abstract hierarchy of formalisms, without any clear sense of what kinds of physical systems, and which of their properties, are at issue.

Stevan Harnad Cognitive Science Laboratory | Laboratoire Cognition et Mouvement Princeton University | URA CNRS 1166 I.B.H.O.P. 221 Nassau Street | Universite d'Aix Marseille II Princeton NJ 08544-2093 | 13388 Marseille cedex 13, France harnad@princeton.edu | harnad@riluminy.univ-mrs.fr 609-921-7771 | 33-91-66-00-69

From harnad Sat Jun 12 14:38:38 1993 To: sharnad@life.jsc.nasa.gov Subject: Re: Prima facie question

Date: Sat, 12 Jun 93 10:40:46 -0400 From: davism@turing.cs.nyu.edu (Martin Davis)

I haven't read Yee's paper, but it seems to me that he is badly confused.

I've written a pair of technical papers a long time ago on the question precisely of when a TM can be regarded as being a UTM. The technical issue is how to separate the computational work in "decoding" the symbol string representing a TM being simulated from the work of that TM itself.

The conclusion is that any "sufficiently powerful" TM can be regarded as a UTM. Martin Davis

----

Date: Mon, 14 Jun 93 17:14 BST From: ronc@cogs.susx.ac.uk (Ron Chrisley)

Hello, Stevan.

A warning. I have not read Yee's paper. But some things you said while discussing it prompted me to make this quick comment (a similar point appears in my Minds & Machines draft, though perhaps not as explicitly as it should).

Date: Fri, 11 Jun 93 22:25:30 EDT From: "Stevan Harnad"

That all sounds fine, but one still needs the answers to a few questions:

(1) What kind of system is NOT a TM then, in this sense? Is a bridge, a furnace, a plane, a protein, an organism, an atom, a solar system? a brain? the universe?

(2) If the answer to all of the above is that they are all TMs, then OF COURSE systems with minds are TMs too, but then what does this tell us about the mind that isn't true about everything else under the sun (including the sun) as well? Saying it was a TM was supposed to tell us something more than that it was a physical system! We ALL accepted that in the first place. The question was, what KIND of physical system. If "TMs" does not pick out a subset, it's not informative, just as it would not be informative to tell us that the brain, like everything else, obeys differential equations.

I maintain that even if everything can be understood to be a TM (i.e., for everything, there is at least one TM description which applies to it) this does not make the idea that "cognition is computation" vacuous, on a suitably reasonable reading of that motto. The reasonable reading is not "anything that has a computational description is a cognizer"; that reading, combined with ubiquity of computation, would indeed lead to panpsychism, which many are loathe to accept. Rather, the reasonble reading is something like "anything that has a TM description which falls in class C is a cognizer", for some *natural (i.e., non-disjunctive and non-question-begging)* class C of TM's. The claim that cognition is computation does not mean that *all* computation is cogntion.

Even if you don't spell out what C is, if you subscribe to the motto, then you are saying something more than "cognition is physically realizable". You are committing yourself to the claim that there will be a natural division between cognizers and non-cognizers using the concepts of computational theory. Which is a substantial claim.

Of course you, Stevan, were well aware of this, but I wanted to clear up a possible misreading.

Ronald L. Chrisley (ronc@cogs.susx.ac.uk) School of Cognitive & Computing Sciences University of Sussex, Falmer, Brighton, BN1 9QH, UK

Date: Sun, 13 Jun 93 13:02:38 -0400 From: yee@envy.cs.umass.edu Subject: SG-list: Reply to M. Davis

Davis expresses grave doubts about whatever it is he imagines I have said about TM's and UTM's. Unfortunately, he cannot explain very well what the alleged problems with the paper are. Nevertheless, I'll take a stab at speculating on the possible relevance of his comments. I will try to briefly spell out my position on universal and non-universal Turing machines.

If the brain is essentially a TM, then surely it is an *online* one, meaning that it must respond in "real-time," producing sequences of outputs that correspond to sequences of inputs. Let us discretize time at some fine level of granularity, t = 0, 1, ..., and suppose that the brain at time t is the online machine Mt, where M0 is some TM; is the input-output pair at time t, and for t > 0,

Mt(X) = M0(, , ..., ; X).

In other words, the output of machine Mt may be a function of its entire input-output history from time 0 to t-1 (in addition to input xt). Under this view, then, instead of corresponding to a fixed TM, the brain may correspond to an evolving sequence of related but possibly slightly different TM's, which may be partially determined by specific input-output experiences. It seems extremely unlikely that the sequence of machines {Mt} would correspond to "the brain's being a UTM" in any interesting sense.

Now, Davis has proved that if a given machine Mt is "sufficiently powerful" (in a precisely defined sense), then, as I understand it, there exist recursive *encoding functions* (which are not themselves universal), under which Mt could be interpreted as a UTM. Denote such encoding functions by Et. Given the undoubted complexity of the brain, it might be that for each Mt there would exist suitable encodings Et, which would then make it possible to interpret each Mt as a UTM.

Would this make the brain essentially a UTM? Given input xt at time t, the transition of machine Mt into M(t+1):

Mt(xt) = yt ---> M(t+1)(X) = Mt(; X),

means that even if it were possible to interpret M(t+1) as a UTM, doing so might require new encoding functions E(t+1), not equal to Et. Hence, even if at each instant there were *some* interpretation under which the current state of the brain would correspond to a UTM, such interpretations might need to be determined anew at each time step.

My reaction is that viewing the brain as a UTM would only be interesting if there were *fixed* encoding functions under which for all t, Mt would be interpretable as a UTM. Such would be the case, for example, if M0 were the description of a common programmable computer because for all t, Mt would equal M(t+1), i.e., a computer's performance in running a program on an input is unaffected by any previous programs and inputs that it might have run. In contrast to this, if viewing the brain as a UTM required perpetual re-interpretation, then it would not be a coherent UTM but a sequence of *accidental UTM's* that would happen to arise due to the complexity of the machines {Mt}. In such a case, it would be more sensible to find a stable and continuous non-UTM account of brain processes (e.g., the presumably non-UTM machines {Mt}).

My apologies for this somewhat involved response to Davis's simple remarks. I must admit that the paper itself does not explain these points in such detail. It also does not address Davis's results concerning UTM's. Perhaps it should. Doing so might make certain statements in the paper more precise, but I doubt very much that it would affect the substance of what is said.

Richard Yee

Date: Mon, 14 Jun 93 18:23:54 -0400 From: yee@envy.cs.umass.edu

Please note: for the next two to three weeks, I will be mostly unable to participate in these discussions. ry

Based on reading the final pages of my paper, Stevan Harnad summarizes one of its main assertions:

> Date: Fri, 11 Jun 93 22:25:30 EDT > From: Stevan Harnad > Subject: Prima facie questions > > Comments on Yee's Paper: > > I could only print out page 16 and upward of Yee's paper, but ... > ... Yee seems to be saying that > only UTMs (Universal Turing Machines, e.g. digital computers) are > symbol-manipulators, and hence open to the objections against pure > symbol manipulation. TMs (Turing Machines) are not. UTMs merely > SIMULATE TMs, which are in turn OTHER kinds of machines that are NOT > just symbol manipulators. Systems with minds are TMs, not UTMs.

This is essentially correct, but I would add a few comments:

(a) Obviously UTM's are TM's, hence *some* TM's perform a certain type of formal symbol processing.

(b) Technically, I only claim that NO ONE HAS PROVEN that all TM's are formal processors---a necessary result for establishing a Chinese room-like argument against all computational accounts of mind. However, I also try to show why such a proof does not exist.

(c) The main thesis of the paper is that the philosophy of mind needs to focus its attention on the complete TM model. Analyses of UTM's and symbol processing are used to show how failing to do so has led to problems, e.g., the Chinese room debate. Analysis of the Godelian debate provides further evidence.

Harnad continues: > That all sounds fine, but one still needs the answers to a few questions: > > [*** PARAPHRASING: > 1. What kind of physical systems would NOT be TM's? > > 2. If all physical systems are TM's, then what is learned by saying > the mind [/brain?] is as well? > > 3. If NOT all systems are TM's, then what exactly is the TM/non-TM > distinction for physical systems, particularly as it regards > minds [/brains]? ***] > > These are the kinds of things you have to be explicit about, otherwise > your reader is lost in a abstract hierarchy of formalisms, without any > clear sense of what kinds of physical systems, and which of their > properties, are at issue.

Harnad's questions focus on the relationship between the physical world and the mathematical theory of TM-computation. I agree that this is an important philosophical issue which should be addressed. However, the purpose of my paper is to examine what the theory of computation says about the Chinese room and Godelian arguments. The theory shows that neither argument can fully refute computationalism because neither fully addresses the TM model. The conclusion is that all sides would be better served by focusing on TM's proper (instead of, e.g., UTM-computers or formal systems).

In contrast, the theory of computation says almost nothing about the physical world. Except for the discreteness and finiteness of TM processing steps, it does not specify how TM's are or are not realized by physical systems. Thus, the questions raised by Harnad, do not lie within the scope of either the theory or the current paper. As Ron Chrisley alluded to in his recent message, such questions *are* currently being addressed in the "what is computation?" discussion.

On the one hand, Harnad's questions raise the concern that computationalism might be *too broad*, being nothing more than a restatement of the view that minds arise from the actions of physical systems: brains. On the other hand, critiques such as the Chinese room and Godelian arguments claim that computation is *too limited*, specifically lacking the ability to process symbols semantically, in the same manner as minds. The concerns of the "too-broad" view are well-founded because TM-computation is quite general indeed, as the Church-Turing thesis atests. This makes the claims of the "too-limited" arguments seem all the more provocative. Furthermore, both the Chinese room (CR) and Godelian arguments are long-standing, widely debated, and widely accepted as definitive critiques of computationalism. Therefore, it is sufficient to refute them alone, without taking on the additional valid concerns raised by Harnad.

The paper tries to provide clear refutations of these two prominent critiques. Hopefully, the refutations remain mostly within the theory of computation, calling for understanding rather than belief in anyone's intuitions. The "what is computation?" issues raised by Harnad, besides being somewhat opposite in orientation, appear to lie mostly beyond the scope of the mathematical theory. Richard Yee

----

Date: Tue, 15 Jun 93 20:37:01 EDT From: "Stevan Harnad"

Date: Tue, 15 Jun 93 18:35 BST From: ronc@cogs.susx.ac.uk (Ron Chrisley)

>
>sh> Date: Mon, 14 Jun 93 14:50:15 EDT
>
>sh> From: "Stevan Harnad"

> rc> I maintain that even if everything can be understood to be a TM (i.e.,
> rc> for everything, there is at least one TM description which applies to
> rc> it) this does not make the idea that "cognition is computation"
> rc> vacuous, on a suitably reasonable reading of that motto. The
> rc> reasonable reading is not "anything that has a computational
> rc> description is a cognizer"; that reading, combined with ubiquity of
> rc> computation, would indeed lead to panpsychism, which many are loathe
> rc> to accept. Rather, the reasonable reading is something like "anything
> rc> that has a TM description which falls in class C is a cognizer", for
> rc> some *natural (i.e., non-disjunctive and non-question-begging)* class
> rc> C of TM's. The claim that cognition is computation does not mean that
> rc> *all* computation is cognition.

> rc> Even if you don't spell out what C is, if you subscribe to the motto,
> rc> then you are saying something more than "cognition is physically
> rc> realizable". You are committing yourself to the claim that there will
> rc> be a natural division between cognizers and non-cognizers using the
> rc> concepts of computational theory. Which is a substantial claim.

>
>sh> Ron, of course I am aware that the hypothesis that All Cognition Is
>
>sh> Computation does not imply that All Computation Is Cognition, but that
>
>sh> still does not help. Sure if the brain and the sun are both doing
>
>sh> computations, they're still doing DIFFERENT computations, and you're
>
>sh> free to call the sun's kind "solar computation" and the brain's kind
>
>sh> "cognitive computation," but then you may as well have said that they
>
>sh> both obey differential equations (different ones), and the sun obeys
>
>sh> the solar kind and the brain the cognitive kind. That's obvious a
>
>sh> priori and does not help us one bit either.

Yes, that’s why I said C has to be a *natural* and *non-question-begging* class of computation. Simply *defining* what the brain does as "cognitive computation" is not going to get one anywhere. One has to show that there is a class, naturally expressible in terms of computational concepts, that includes brains and all other cognizing physical systems, but leaves out stars, stones, etc. Only then will one be justified in claiming "cognition is computation" in any non-vacuous sense. If the best one can do, when using computational concepts, is find some wildly disjunctive statement of all the systems that are cognizers, then that would suggest that cognition is *not* computation. So the claim is not vacuous, but contentious.

>
\>sh> In fact, computationalism was supposed to show us the DIFFERENCE
>
\>sh> between physical systems like the sun and physical systems like the
>
\>sh> brain (or other cognitive systems, if any), and that difference was
>
\>sh> supposed to be that the brain’s (and other cognitive systems’)
>
\>sh> cognitive function, UNLIKE the sun’s solar function, was
>
\>sh> implementation-independent -- i.e., differential-equation-independent
>
\>sh> -- because cognition really WAS just (a kind of) computation; hence
>
\>sh> every implementation of that computation would be cognitive, i.e.
>
\>sh> MENTAL (I have no interest whatsoever in NON-mental "cognition," because
>
\>sh> that opens the doors to a completely empty name-game in which we call
>
\>sh> anything we like "cognitive").

That might have been how many people thought computation was relevant to cognitive science, but then one can take what I say here to be a different proposal. I think both the sun and the brain can be looked at as performing some computation. So what’s special about cognition is not that it is realized in a system that can be looked at computationally.

Nor is the multiple realizability feature significant here. Once one admits that there is a computational description of what the sun is doing, then one ipso facto admits that in some sense, what the sun is doing is multiply realizable too. So that’s not what is so computationally special about cognition.

What’s special is this: there is no *natural*, *non-wildly-disjunctive* way to distinguish white dwarf stars from red giant stars by appealing to the computational systems they instantiate. The "cognition is computation claim", however, is claiming that there *is* a natural way to distinguish cognizing systems from non-cognizing ones. And that natural way is not using the concepts of neurophysiology, or molecular chemistry, but the concepts of computational theory. This is a substantive claim; it could very well be false. It certainly isn’t vacuously true (note that it is not true for the case of stars); and its bite is not threatened even if everything has a computational characterization.

Now if everything can have *every* computational characterization, then the claim might be in danger of being content-free. But that's why I took the time to rebutt Putnam's (and in some sense, Searle's) arguments for that universal-realization view.

>
>sh> This is where Searle's argument came in. showing that he could himself
>
>sh> become an implementation of that "right" cognitive system (the
>
>sh> T2-passing symbol cruncher, totally indistinguishable from a life-long
>
>sh> pen-pal), but without having the requisite cognitive (= mental) state,
>
>sh> namely, understanding what the symbols were ABOUT.

I agree that Searle was trying to show that the "cognition is computation" claim is false. But his argument applies (although I don't feel it succeeds) to my construal of the "C is C" claim. He was trying to show that there was no such class of computation that characterizes cognizers, since he could instantiate one of the computations in any proposed class and not have the requisite cognitive properties.

>
>sh> Now the implementation-independence of the symbolic level of description
>
>sh> is clearly essential to the success of Searle's argument, but it seems
>
>sh> to me that it's equally essential for a SUBSTANTIVE version of the
>
>sh> "Cognition Is Computation" hypothesis (so-called "computationalism").
>
>sh> It is not panpsychism that would make the hypothesis vacuous if
>
>sh> everything were computation; it would be the failure of "computationalism"
>
>sh> to have picked out a natural kind (your "C").

As computational theory stands now, it is pretty implementation-independent. Such a purely formal theory may or may not (I go with the latter) be the best account of computation. But even if our notion of computation were to change to a more "implementation-dependent" notion (although I suspect we wouldn't think of it as more "implementation-dependent" once we accepted it), I don't see why the "C is C" claim would be in danger of vacuity. It would be even stronger than before, right? But perhaps you just meant that it would be false, since it would rule out cognizers made of different stuff? That's just a guess that the "C is C" claim is false: that the natural kinds of computation do not line up with cognition. That's not an argument.

>
>sh> For we could easily have had a "computationalist" thesis for solar heat
>
>sh> too, something that said that heat is really just a computational
>
>sh> property; the only thing that (fortunately) BLOCKS that, is the fact
>
>sh> that heat (reminder: that stuff that makes real ice melt) is NOT
>
>sh> implementation-independent, and hence a computational sun, a "virtual
>
>sh> sun" is not really hot.

Right, but the computation that any particular hot thing realizes *is* implementation-independent. The question is whether that computation is central to the phenomenon of heat, or whether it is accidental, irrelevant to the thing *qua* hot thing.

If you want to avoid begging the question "is heat computation?", then you have to allow that heat *might* be implementation-independent. Then you notice that there is no natural computational class into which all hot things fall, so you reject the notion of "heat is computation" and thereby the prospects for implementation-independence.

>
>sh> In contrast, computationalism was supposed to have picked out a natural
>
>sh> kind, C, in the case of cognition: UNLIKE HEAT ("thermal states"),
>
>sh> mental states were supposed to be implementation-independent symbolic
>
>sh> states. (Pylyshyn, for example, refers to this natural kind C as
>
>sh> functions that transpire above the level of the virtual architecture,
>
>sh> rather than at the level of the physical hardware of the
>
>sh> implementation; THAT's what's supposed to distinguish the cognitive
>
>sh> from the ordinary physical).

No. *All* formal computations are implementation-independent. So it cannot be implementation-independence that distinguishes the cognitive from the non-cognitive (now you see why I reminded us all that "all cognition is computation" does not mean "all computation is cognition"). There are many other phenomena whose best scientific account is on a computational level of abstraction (what goes on in IBM's, etc.; perhaps some economic phenomena, et al). So the fact that a phenomenon is best accounted for on the computational (implementation-independent) level does not mean that it is cognitive. It means that it is *computational*. The big claim is that cognition is a computational phenomenon in this sense.

>
>sh> And it was on THIS VERY VIRTUE, the one
>
>sh> that made computationalism nonvacuous as a hypothesis, that it
>
>sh> foundered (because of Searle's argument, and the symbol grounding
>
>sh> problem).

I accept that "implementation-dependent" (better: not purely formal) notions of computation are probably the way to go (I've argued as much on this list before), but I don't feel that Searle compels me to do this. The sociology of computer science certainly hasn't worked that way. It just so happens that embedded and embodied notions are essential to understand normal, ordinary computational systems like token rings, CPU's hooked to robot arms, etc. But there is a disagreement here: if computational theory goes embedded (rejects implementation-independence), that doesn't mean it is vacuous; just the opposite! It makes its range of application even more restricted.

My original claim (the negation of the one you made in your initial response to Yee) was that *even if* everything has a computational characterization, that does not make the "computation is cognition" claim vacuous. I have given the reasons above. Now if we have an implementation-dependent computational theory, that does not mean that not everything will have a computational characterization. It could just mean that tokens that were of the same computational type in the formal theory are now of distinct types in the embedded theory. Nevertheless, despite such ubiquity of computation, there might still be a natural computational kind which includes just those things which are cognizers. Or there might not.

Ronald L. Chrisley (ronc@cogs.susx.ac.uk) School of Cognitive & Computing Sciences University of Sussex

----

From: harnad@clarity.princeton.edu (Stevan Harnad) Date: Tue Jul 6 21:06:54 EDT 1993

ON IMPLEMENTATION-DEPENDENCE AND COMPUTATION-INDEPENDENCE

Ron Chrisley suggests that the thesis that "Cognition is Computation" is nonvacuous even if every physical process is computation, as long as COGNITIVE computation can be shown to be a special KIND of computation. (He does not, of course, suggest what that special kind of computation might actually be, for here he is only trying to establish that such a thesis is tenable.)

Ron does mention (and to a certain extent equivocates on) one difference that may indeed distinguish two different kinds of computation: the "implementation-DEpendent" kind (again, not defined or described, just alluded to) and the usual, implementation-INdependent kind. Ron thinks cognition may turn out to be implementation-DEpendent cognition.

In writing about physical systems that science has NOT so far found it useful to describe, study or explain computationally (the sun, for example), Ron notes that they are nevertheless computational systems, and in some sense "multiply realizable" (I'm not sure whether he means this to be synonymous with implementation-independent -- I don't think there's any cat that can only be skinned ONE way, but I don't think that's quite what's ordinarily meant by "implementation-independence" in the computational context, otherwise that term too risks becoming so general as to become vacuous.)

I'll lay my own cards on the table, though: The only implementation-independence *I* think is relevant here is the kind that a computer program has from the computer that is implementing it. Is that the kind Ron thinks the sun has? (I mean the sun in our solar system here, not the workstation by that name, of course!) If so, then a computer simulation of the sun -- a physical symbol system implementing the sun's COMPUTATIONAL description, in other words, a computer running the sun's computer program and hence systematically interpretable as the sun, a virtual sun -- would have to be hot (VERY hot, and not just symbols systematically interpretable as if they were hot). We don't even need Searle to see that THAT's not likely to happen (because you can FEEL the absence of heat, but you can't be so sure about the absence of mind). So whatever kind of "implementation-independence" the sun may have, it's not the kind we need here, for the purposes of the "cognition is computation" thesis.

So suppose we give up on that kind of software/hardware implementation-independence and settle for "implementation-DEpendent" computation -- whatever that is, for it sounds as if spelling out the nature of that dependence will turn out to be as essential to a description of such a "computational" system as the computational description itself. Indeed, it sounds as if the dependence-story, unlike the computation-story, will turn out to be mostly physics in that case. I mean, I suppose flying is an implementation-dependent sort of computation too, and that a plane is, in a sense, just a highly implementation-dependent computer. The only trouble is that all the RELEVANT facts about planes and flying are in the physics (aeronautical engineering, actually) of that dependency, rather than the computation!

So if we open up the Pandora's box of implementation-dependence, is there not the risk that the "Cognition is (implementation-dependent) Computation" thesis would suffer the same fate as a "Flying is (implementation-dependent) Computation" thesis?

Now to the play-by-play:

rc> Date: Tue, 15 Jun 93 18:35 BST rc> From: ronc@cogs.susx.ac.uk (Ron Chrisley)

rc> I said C has to be a *natural* and rc> *non-question-begging* class of computation. Simply *defining* what rc> the brain does as "cognitive computation" is not going to get one rc> anywhere. One has to show that there is a class, naturally rc> expressible in terms of computational concepts, that includes brains rc> and all other cognizing physical systems, but leaves out stars, stones, rc> etc. Only then will one be justified in claiming "cognition is rc> computation" in any non-vacuous sense. If the best one can do, when rc> using computational concepts, is to find some wildly disjunctive rc> statement of all the systems that are cognizers, then that would rc> suggest that cognition is *not* computation. So the claim is not rc> vacuous, but contentious.

It's not coincidental that the "Is Cognition Computation?" and the "What is Computation?" discussion threads are linked, because it's critical to get it straight what it is that we are affirming or denying when we say Cognition Is/Isn't Computation. I was content to assume that what was at issue was implementation-independent symbol manipulation, but now, with the introduction of TMs (Turing Machines) in place of UTMs (Universal Turing Machines) it's being suggested that that's not the issue after all. It seems to me that although your burden is not to actually produce the

RIGHT theory of what is special about that subset of computation that is cognitive, you do have to give us some idea of the KIND of thing it might be.

So let's look more closely at your implementation-dependent TM theory of mind. Here's an important question: Would the UTM simulation of the right TM (the one that had mental states) have mental states? If so, we're back to Searle. If not, I'd like to know why not, since computational equivalence is supposed to be the pertinent INVARIANT that holds all these computational descriptions together. I mean, without the computational equivalence, isn't it back to physics again?

>
\>sh> In fact, computationalism was supposed to show us the DIFFERENCE
>
\>sh> between physical systems like the sun and physical systems like the
>
\>sh> brain (or other cognitive systems, if any), and that difference was
>
\>sh> supposed to be that the brain's (and other cognitive systems')
>
\>sh> cognitive function, UNLIKE the sun's solar function, was
>
\>sh> implementation-independent -- i.e., differential-equation-independent
>
\>sh> -- because cognition really WAS just (a kind of) computation; hence
>
\>sh> every implementation of that computation would be cognitive, i.e.
>
\>sh> MENTAL

rc> That might have been how many people thought computation was relevant rc> to cognitive science, but then one can take what I say here to be a rc> different proposal. I think both the sun and the brain can be looked rc> at as performing some computation. So what's special about cognition rc> is not that it is realized in a system that can be looked at rc> computationally.

(I don't mean to pick on your phraseology, but that last sentence sounds like a denial of the Cog=Comp thesis right there...) But of course you are here adverting to the fact that it's going to turn out to be a special KIND of computation. Can you be more specific, perhaps give examples of other systems that have been taken to be natural kinds because they turned out to be special kinds of computational systems? And can you suggest what lines the specialness might take in the case of cognition?

rc> Nor is the multiple realizability feature significant here. Once one rc> admits that there is a computational description of what the sun is rc> doing, then one ipso facto admits that in some sense, what the sun is rc> doing is multiply realizable too. So that's not what is so rc> computationally special about cognition.

As I said, multiple-realizability is not quite the same as implementation-independence. There are, for example, many different ways to transduce light, some natural (the vertebrate retinal cone or the invertebrate omatidium), some artificial (as in a bank-door's photosensitive cell), but NONE of them are computational, nor constitute a natural family of computationally equivalent systems -- or if they do, the computational story is trivial. It's the physics of light-transduction that's relevant.

By way of contrast, all the things you can reconfigure a computer to DO by changing its software DO share interesting properties, and the properties are computational ones: The same software can be run on radically different forms of hardware yet it would still be performing the same computation. THAT was the kind of multiple-realizability that I THOUGHT was relevant to what computation is and what Cog=Comp claims.

By way of contrast, note that the number of ways you can reconfigure a UTM like a digital computer to implement different programs does NOT include a way to reconfigure it into an optical tranducer, a plane, or a sun. For that, the "reconfiguring" would have to be more radical than merely computational: It would have to be physical. (And that's why I think implementation-DEpendent "computation" is a non-starter.)

rc> What's special is this: there is no *natural*, *non-wildly-disjunctive* rc> way to distinguish white dwarf stars from red giant stars by appealing rc> to the computational systems they instantiate. The "cognition is rc> computation claim", however, is claiming that there *is* a natural way rc> to distinguish cognizing systems from non-cognizing ones. And that rc> natural way is not using the concepts of neurophysiology, or molecular rc> chemistry, but the concepts of computational theory. This is a rc> substantive claim; it could very well be false. It certainly isn't rc> vacuously true (note that it is not true for the case of stars); and rc> its bite is not threatened even if everything has a computational rc> characterization.

But, as I said, I UNDERSTOOD the claim when I took computation to be implementation-independent symbol-manipulation. But with implementation-DEpendent TMs I no longer even know what's at issue... "Not wildly disjunctive" just isn't a positive enough characterization to give me an inkling. Do you have examples, or (relevant) analogies?

Let me put it even more simply: It's clear that some subset of computer programs is the subset that can do, say, addition. Let's suppose that this subset is "not wildly disjunctive." It is then an equivalence class, of which we can say, with confidence, that every implementation of those computer programs will be doing addition. Now all you need is a similar story to be told about thinking: Find the right ("nondisjunctive") subset of computer programs, and then every implementation of them will be thinking. But now you seem to be saying that NOT every implementation of them will be thinking, because the programs are implementation DEpendent. So what does that leave of the claim that there is a nontrivial COMPUTATIONAL equivalence there to speak of at all?

Remember, if we go back to the sun, the scientific story there is thermodynamics, electromagnetism, etc. It's not in any interesting sense a computational story. Solar physics is not a branch of computer science. No one is espousing a "Solar Dynamics = Computation" hypothesis. All of physics is (approximately) COMPUTABLE, but that does not mean that physical processes are COMPUTATIONAL. And as far as I can tell, the most direct CARRIER of that dissociation is the fact that physics is not implementation-independent. So computational equivalence has a hollow ring to it when you are trying to explain the physics. The burden is to show why exactly the same thing is not true when it comes to explaining thinking.

rc> Now if everything can have *every* computational characterization, rc> then the claim might be in danger of being content-free. But that's rc> why I took the time to rebut Putnam's (and in some sense, Searle's) rc> arguments for that universal-realization view.

As it happens, I never accepted the "everything is EVERY computation" view (for the cryptographic reasons I adduced earlier in this discussion). But I think "everything is SOME [implementation-independent OR implementation-dependent] computation" is just as empty and unhelpful as a basis for a cognition-specific thesis, for that's just the Church-Turing Thesis, which is a formal thesis about what "computation" is and has NOTHING to do with mental states or what they are or aren't.

rc> I agree that Searle was trying to show that the "cognition is rc> computation" claim is false. But his argument applies (although I rc> don't feel it succeeds) to my construal of the "C is C" claim. He was rc> trying to show that there was no such class of computation that rc> characterizes cognizers, since he could instantiate one of the rc> computations in any proposed class and not have the requisite rc> cognitive properties.

Yes, but Searle's argument works only for computation construed as implementation-INdependent symbol manipulation. If some other sense of computation is at issue here, his argument may well fail, but then I don't know what would be at issue in its place.

Indeed, Searle's argument fails immediately if anyone wants to say (as, say, Pat Hayes does): Cognition has to be implemented the "right way" and Searle's implementation is not the right one. But to save THAT from amounting to just special pleading in the same sense that, say, a "wild disjunction" would be, one has to face the problem of how to distinguish the "right" from the "wrong" implementation without shifting the scientific substance of the explanation of cognition to the implementational details rather than the computation!

Again, to put it ever so briefly: Implementation-DEpendent "computation" would indeed be immune to Searle, but look at the price: Cognition is now not just the right computatation, but the right implementation of that computation -- and then the rest is just an arbitrary squabble about proportions (computations/physics).

>
\>sh> Now the implementation-independence of the symbolic level of description
>
\>sh> is clearly essential to the success of Searle's argument, but it seems
>
\>sh> to me that it's equally essential for a SUBSTANTIVE version of the
>
\>sh> "Cognition Is Computation" hypothesis (so-called "computationalism").
>
\>sh> It is not panpsychism that would make the hypothesis vacuous if
>
\>sh> everything were computation; it would be the failure of
>
\>sh> "computationalism" to have picked out a natural kind (your "C").

rc> As computational theory stands now, it is pretty rc> implementation-independent. Such a purely formal theory may or may rc> not (I go with the latter) be the best account of computation.

It’s at this point that I sense some equivocation. We are now to envision not only a still unspecified "special" KIND of computation that is peculiar to (and sufficient for) mental states, but we must also imagine a new SENSE of computation, no longer the old implementation-independent kind on which the whole formal theory was built. My grip on this inchoate "computation" is loosening by the minute...

rc> But even if our notion of computation were to change to a more rc> "implementation-dependent" notion (although I suspect we wouldn’t rc> think of it as more "implementation-dependent" once we accepted it), I rc> don’t see why the "C is C" claim would be in danger of vacuity. It rc> would be even stronger than before, right? But perhaps you just meant rc> that it would be false, since it would rule out cognizers made of rc> different stuff? That’s just a guess that the "C is C" claim is rc> false: that the natural kinds of computation do not line up with rc> cognition. That’s not an argument.

I couldn’t be suggesting that such a claim was false, since, as I said, I’ve lost my grip on what the claim is claiming!

"Stuff" had nothing to do with the old Cog=Comp thesis, since it was implementation-independent. And I could quite consistently manage to be an anticomputationalist toward this old form of computationalism (because of the symbol grounding problem) without for a minute denying that minds could be realized in multiple ways (just as optical transducers can); in fact, that’s what my Total Turing Test (T3) banks on. But SYNTHETIC alternative realizations (made out of different, but T3-equivalent stuff) are not the same as VIRTUAL alternative realizations, which is what a pure symbol system would be. Besides, a symbol-cruncher alone could not pass T3 because it lacks sensorimotor transducers -- significantly, the only part of a robot that you can’t have a virtual stand-in for.

But never mind that: Why would a Cog=Comp thesis involving "implementation-dependent computation" be stronger than one involving implementation-independent computation? It sounds more like a weaker one, for, as I keep hinting, there is the risk that in that case the relevant functions are in the DEPENDENCY, i.e., in the physics (e.g., the transduction) rather than the computation.

>
>sh> For we could easily have had a "computationalist" thesis for solar heat
>
>sh> too, something that said that heat is really just a computational
>
>sh> property; the only thing that (fortunately) BLOCKS that, is the fact
>
>sh> that heat (reminder: that stuff that makes real ice melt) is NOT
>
>sh> implementation-independent, and hence a computational sun, a "virtual
>
>sh> sun" is not really hot.

rc> Right, but the computation that any particular hot thing realizes *is* rc> implementation-independent. The question is whether that computation rc> is central to the phenomenon of heat, or whether it is accidental, rc> irrelevant to the thing *qua* hot thing.

Well, apart from the fact that thermodynamics, the science of heat, shall we say, is not accustomed to thinking of itself as a computational science, surely the ESSENTIAL thing about heat is whatever "being hot" is, and that's exactly what virtual heat lacks. If this is not obvious, think of a virtual plane in a virtual world: It may be computationally equivalent to a real plane, but it can't fly! I wouldn't describe the computational equivalence between the real and virtual plane as accidental, just as insufficient -- if what we wanted was something that FLEW!

And EXACTLY the same is true in the case of wanting something that really has mental states -- at least with the old candidate: implementation-independent symbol manipulation (which could yield only a VIRTUAL mind, not a real one). But I have no idea how to get a grip on an implementation-DEPENDENT candidate. I mean, what am I to suppose as the TM in question? There's (1) real me. I have real mental states. There's (2) virtual "me" in a virtual world in the pure symbol cruncher. It's computationally equivalent to me, but for Searlean reasons and because of the symbol grounding problem, I don't belive for a minute that it has a mind.

But that's not what's at issue here. We should now think of a third entity: (3) A TM performing implementation-DEpendent computations. I can't imagine what to imagine! If it's a robot that's T3-indistinguishable, I'm already ready to accept it as a cognizing cousin, with or without a computational story (it could all be a transducer story -- or even a HEAT story, for that matter, in which case real heat could be essential in BOTH cases). But what am I to imagine wanting to DENY here, if I wanted to deny this new form of computationalism, with TMs and implementation-DEpendence instead of UTMs and implementation-INdependence?

rc> If you want to avoid begging the question "is heat computation?", then rc> you have to allow that heat *might* be implementation-independent. rc> Then you notice that there is no natural computational class into rc> which all hot things fall, so you reject the notion of "heat is rc> computation" and thereby the prospects for rc> implementation-independence.

Ron, you've completely lost me. There's no sense of heat that I can conjure up in which heat is computation, no matter how many ways it can be realized. (Again: multiple-realizability is not the same as implementation-independence.) I have no problem with synthetic heat, but that still does not help me see heat as computation (the real problem is VIRTUAL heat). And even if there is a nice, crisp ("nondisjunctive") set of unique computational invariants that characterize hot things and no others, I still don't see what it would mean to say that heat was computation -- except if EVERY implementation of the heat program were hot -- which is decidedly not true (because virtual heat is not hot). (Also, in the above passage it sounds as if you are conceding that computationality DOES call for implementation-INdependence after all.)

>
\>sh> In contrast, computationalism was supposed to have picked out a natural
>
\>sh> kind, C, in the case of cognition: UNLIKE HEAT ("thermal states"),
>
\>sh> mental states were supposed to be implementation-independent symbolic
>

>sh> states. (Pylyshyn, for example, refers to this natural kind C as
>
>sh> functions that transpire above the level of the virtual architecture,
>
>sh> rather than at the level of the physical hardware of the
>
>sh> implementation; THAT's what's supposed to distinguish the cognitive
>
>sh> from the ordinary physical).

rc> No. *All* formal computations are implementation-independent.

Again, there seems to be some equivocation here. I thought you had said you thought that cognition might be an implementation-DEpendent form of computation earlier. Or is there now "formal computation" and "computation simpliciter" to worry about? Entities seem to be multiplying and it sounds like it's all in the service of saving an increasingly vague if not vacuous thesis...

rc> So it rc> cannot be implementation-independence that distinguishes the cognitive rc> from the non-cognitive (now you see why I reminded us all that "all rc> cognition is computation" does not mean "all computation is rc> cognition"). There are many other phenomena whose best scientific rc> account is on a computational level of abstraction (what goes on in rc> IBM's, etc.; perhaps some economic phenomena, et al). So the fact rc> that a phenomenon is best accounted for on the computational rc> (implementation-independent) level does not mean that it is cognitive. rc> It means that it is *computational*. The big claim is that cognition rc> is a computational phenomenon in this sense.

No, just as I have indictated that I of course realize that "all cog is comp" does not imply "all comp is cog," I don't think that the emphasis on the implementation-independence of cognition was enough to uniquely characterize cognition, for there are of course plenty of other implementation-independent computational phenomena. It had more of a necessary- than a sufficient-condition flavor -- though of course necessity was not at issue at all. What Pylyshyn was saying was that mental states were unlike other kinds of physical states, and were like other kinds of computational states (including nonmental ones) in being IMPLEMENTATION-INDEPENDENT. That wasn't SUFFICIENT to make every computation mental, but it was sufficient to distance cognitive science from certain other forms of physicalism, in particular, the kind that looked for a hardware-level explanation of the mind.

Now I had asked for examples. IBM is a bad one, because it's a classical (approxiation to a) UTM. Economic phenomena are bad examples too, for of course we can model economic phenomena (just as we can model solar systems), and all that means is that we can predict and explain them computationally. I would have conceded at once that you could predict and explain people and their thoughts computationally too. I just don't think the computational oracle actually THINKS in so doing, any more than the planetary oracle moves (or the virtual plane flies or the virtual sun is hot). "Economics" is such an abstract entity that I would not know what to do with that; it's not a concrete entity like a person or a set of planets. If you make it one, if you say you want to model "society" computationally, then I'll say it's the same as the solar system oracle. There's no people in the one, no planets in the other, because such things are not implementation-independent.

>
>sh> And it was on THIS VERY VIRTUE, the one
>
>sh> that made computationalism nonvacuous as a hypothesis, that it
>
>sh> foundered (because of Searle's argument, and the symbol grounding
>
>sh> problem).

rc> I accept that "implementation-dependent" (better: not purely formal) rc> notions of computation are probably the way to go (I've argued as much rc> on this list before), but I don't feel that Searle compels me to do rc> this.

Fine, but I have no idea what I am committing myself to or denying if I accept or reject this new, "nonformal" form of computationalism. I doubt if Searle would know either. He has, for example, never denied the possibility of synthetic brains -- as long as they have the relevant "causal powers" of the real brain. A pure symbol-cruncher, he has shown, does NOT have the relevant causal powers. So if you tell him that all the systems that DO have that causal power are computationally equivalent, he'll just shrug, and say, fine, so they are, just as, perhaps, all stars are computationally equivalent in some way. The relevant thing is that they have the right causal powers AND IT'S NOT JUST COMPUTATION, otherwise the virtual version -- which is, don't forget, likewise computationally equivalent -- would have the causal powers too, and it doesn't. Now that you've disavowed UTMs, pure formal sytnax and implementation-independence, this does not bother you, Ron; but just what, exactly, does it leave you with?

rc> The sociology of computer science certainly hasn't worked that rc> way. It just so happens that embedded and embodied notions are rc> essential to understand normal, ordinary computational systems like rc> token rings, CPU's hooked to robot arms, etc. But there is a rc> disagreement here: if computational theory goes embedded (rejects rc> implementation-independence), that doesn't mean it is vacuous; just rc> the opposite! It makes its range of application even more restricted.

Symbol grounding is not just "embedded" symbol crunchers with trivial add-on peripherals; but never mind. I do agree that T3 exerts constraints on ALL models (whether computational or, say, analog-transductive), constraints that T2, imagination, and virtual worlds alone do not. I've already declared that I'm ready to confer personhood on the winning T3 candidate, no matter WHAT's going on inside it. The only thing that has been ruled out, as far as I'm concerned, is a pure symbol cruncher.

But since you never advocated a pure symbol cruncher, you would have to say that Searle is right in advocating a full neuromolecular (T4) understanding of the brain, because the brain, like everything else, is a computational system, and if "C is C" is right, Searle will end up converging on the very same computational theory everyone else does.

My own guess is that going "embedded" or "implementation-dependent" amounts to conceding that the cognition is in the physics rather than just the computation. What shape that physics actually ends up taking -- whether it is just a matter of hooking up the right peripherals to a symbol cruncher in order to make the mental lights go on or (as I suspect) there's rather more to it than that -- is beside the point. The logical implication stands that without the (shall we call it) "computation-independent" physics -- the RIGHT (non-wildly disjunctive) physics -- there is no cognition, even if the computation is "right."

rc> My original claim (the negation of the one you made in your initial rc> response to Yee) was that *even if* everything has a computational rc> characterization, that does not make the "computation is cognition" rc> claim vacuous. I have given the reasons above. Now if we have an rc> implementation-dependent computational theory, that does not mean that rc> not everything will have a computational characterization. It could rc> just mean that tokens that were of the same computational type in the rc> formal theory are now of distinct types in the embedded theory. rc> Nevertheless, despite such ubiquity of computation, there might still rc> be a natural computational kind which includes just those things which rc> are cognizers. Or there might not. Ron Chrisley

I don't know about types and tokens, but if there are two physical systems that are both implementations of the very same formal (computational) system and one of them is "right" and the other one is "wrong," then it sounds as if the formal (computational) story is either incorrect or incomplete. To resolve the ambiguity inherent in computationalism it is therefore not enough to point out, as Ron has done, that the "Cognition is Computation" thesis just claims that "Cognition is a KIND of Computation"; for what it really claims is that "Cognition is JUST a Kind of Cognition." And it's that "JUST" that I think implementation-DEpendence then gives up. But once that's given up, of course, anything goes, including that the computational aspects of cognition, having already been conceded to be partial, turn out to be minimal or even irrelevant...

Stevan Harnad

Date: Thu, 8 Jul 93 11:02:32 EDT From: "Stevan Harnad"

rk> From: Robert.Kentridge@durham.ac.uk rk> Subject: Re: "Is Cognition Computation?" rk> Date: Wed, 7 Jul 1993 14:50:10 +0100 (BST) rk> rk> As I'm very interested in classifying physical systems according rk> to their intrinsic computational properties I thought I'd offer a few rk> comments on your recent exchange with Ron Chrisley. rk> rk> Given some criteria as to what constitutes a good computational rk> description of systems (for example, ones in which the graph rk> indeterminacy of the machine describing the computation is minimized) rk> it is easy (in principle, although quite an effort in practice!) to rk> produce computational descriptions of physical systems (e.g. the sun, rk> a brain, a neuron, a neural network model). Crutchfield and Young rk> describe an algorithm to do just this; from it we can produce rk> stochastic symbolic computational descriptions of any system from rk> which we can make a series of quantified observations over time. The rk> details of the C&Y algorithm are irrelevant here; all we need to rk> concentrate on is the fact that any dynamics can have a symbolic rk> computational description. rk> rk> One problem we face in producing symbolic descriptions of systems is rk> deciding what to measure when we prepare a time-series for our rk> chosen algorithm. If I am producing computational descriptions of rk> stars with the aim of discovering the common computational principles rk> underlying starhood should I measure time series of the luminosity of rk> those stars or should I measure time series of the spatial positions rk> of all of the elementary particles constituting those stars? If I rk> choose the latter course then the computation inherent in luminosity rk> dynamics might emerge as a feature of particle dynamics computation rk> but even so I might not recognize it. Luminosity dynamics may even be rk> of so little predictive power in describing the long term evolution of rk> particle positions that its effects are omitted from the particle rk> dynamics derived description. The problem with implementation rk> independent computation in the context of cognition is that it implies rk> that a system has only one dynamics to measure and that this dynamics rk> underlies cognition. rk> rk> A good reason to worry about the transduction of the external world in rk> cognitive systems is that the nature of this transduction provides us rk> with some clues as to which features of the system from which a rk> computational description might be derived are of functional rk> importance to cognition and which aren't. We might discover some rk> relationship between cognition and computation if we investigate rk> computational descriptions of the dynamics of those features. On the rk> other hand, the intrinsic computation of functionally unimportant rk> features of the system is unlikely to further our understanding of rk> cognition. (If the system in which we believe cognition occurs is the rk> head then my bet would be that studying computational descriptions of rk> hair-growth dynamics won't get us far!!). rk> rk> ps I've now got some data on symbolic machine reconstructions from rk> biologically plausible network models, a tech report and/or preprints rk> should be available soon. rk> rk> Dr. R.W. Kentridge phone: +44 91 374 2621 rk> Psychology Dept., email: robert.kentridge@durham.ac.uk rk> University of Durham, rk> Durham DH1 3LE, U.K.

ON COMPUTATIONAL DESCRIBABILITY VS. ESSENTIAL COMPUTATIONALITY: Computationalism is not just the Church-Turing Thesis

What is at issue in the computationalist thesis that "Cognition IS (just a kind of) Computation" is not whether cognition is DESCRIBABLE by computation. That's conceded at once (by me, Searle, and anyone else who accepts some version of the Church-Turing Thesis). The question is whether it IS just computation. That's why implementation-independence is so critical.

When this is made perfectly explicit:

(1) Thinking IS just (implemented) computation (2) Hence every physical implementation of the right computation will be thinking

then the troubles with this thesis become much clearer (as Searle's Argument and the Symbol Grounding Problem show). In a nutshell, the question becomes: Is a virtual mind really thinking? Is the Universal Turing Machine (UTM), the pure symbol manipulator, that LIKEWISE implements the same computation that describes, say, the brain, THINKING? For if it is not, then Cognition is NOT (just a kind of) Computation.

I think the problem lies squarely with the UNOBSERVABILITY of mental states, which are nevertheless real (see Harnad 1993). That's the only reason this question cannot be settled as trivially as it can in the case of flying and heat. No one would doubt that one could have a full computational DESCRIPTION of a plane or a sun, but no one would dream that the computer implementation of that description (a virtual plane or a virtual sun) actually flew or got hot. Hence no one would say something as absurd as that "Flying (Heating) IS (just a kind of) Computation." Observation alone shows that the critical property is completely missing from the UTM implementation, so it CAN'T be just computation.

With thinking (cognition) we equivocate on this, partly because (a) thinking is unobservable except to the thinker, and partly because we weasel out of even that one by helping ourselves to the possibility of (b) "unconscious thinking" -- which is then unobservable to ANYONE. (The trouble is that, until further notice, unconscious thoughts only occur in the heads of systems that are capable of conscious thoughts, so we are back to (a).) So in my view most of the tenacity of the "Cognition is Computation" thesis derives from the fact that it is not OBVIOUSLY false, as it is in the case of flying and heat, because cognition (or its absence) is unobservable -- to a 3rd person. Yet it IS observable to a 1st person, and this is where Searle's "periscope" comes in.

So, to summarize, computational DESCRIBABILITY is not at issue; essential computationality is. To put it another way, what's at issue is the COMPUTATONALIST Thesis (that C = C), NOT the Church-Turing Thesis that (almost) everything is computationally describable and computationally equivalent to a UTM simulation. It's a mistake to conflate the two theses.

I might also add that differential equations are formulas too. No one doubts that they (and their like) describe all of physics. It is only in a tortured sense that one would want to describe obeying differential equations as "implementing computations," for the whole point of computationalism was to partition phenomena into those that are fully described only by their physics (i.e., differential equations, boundary conditions, etc.,) and those that are fully described by their COMPUTATIONS, independent of their physics. It doesn't help to blur this boundary with a pan-computationalism that even subsumes physics, because then "C = C" REALLY becomes vacuous. (I'm not a physicist, but it seems to me that, irrespective of the success and generality of Crutchfield & Young-style algorithms, physics is not a branch of computer science, even though both use symbols and formulas.)

P.S. Transduction is not just a way of finding out what's functionally important for a computational description of cognition; it IS (part of) what is essential to implementing cognition. (And implementation-independence amounts to a lot more than the fact that hair-growth is irrelevant to brain function.)

Stevan Harnad

Harnad S. (1993) Discussion (passim) In: Bock, G.R. & Marsh, J. (Eds.) Experimental and Theoretical Studies of Consciousness. CIBA Foundation Symposium 174. Chichester: Wiley

----

Date: Fri, 9 Jul 93 10:44:13 EDT From: "Stevan Harnad"

> Date: Thu, 8 Jul 93 17:03:58 BST
> From: Jeff Dalton
>
>
>sh> ON COMPUTATIONAL DESCRIBABILITY VS. ESSENTIAL COMPUTATIONALITY:
>
>sh> Computationalism is not just the Church-Turing Thesis
>
>sh>
>
>sh> What is at issue in the computationalist thesis that "Cognition IS
>
>sh> (just a kind of) Computation" is not whether cognition is DESCRIBABLE
>
>sh> by computation. That's conceded at once (by me, Searle, and anyone else

>
>sh> who accepts some version of the Church-Turing Thesis). The question is
>
>sh> whether it IS just computation. That's why implementation-independence
>
>sh> is so critical.
>
>jd> Why does it follow from the Church-Turing Thesis that cognition
>jd> is describable by computation? For instance, maybe cognition is
>jd> not (completely) describable at all. If so, then _a fortiori_
>jd> it's not describable by computation. This would not show the
>jd> Church-Turing Thesis was false, because all the Thesis would say
>jd> is that if it's describable by computation, then it's descriable
>jd> by the kind of computation performed by Turing machines. If it's
>jd> not descriable by compuation at all, that's no problem for the
>jd> Thesis.

Your point is entirely valid -- and it's because I anticipated it that I said "some version" of the Church-Turing Thesis (C-TT). Of course the standard version of C-TT is just about formal computation and what it is that mathematicians mean by "effective procedure." It is already an extension of C-TT to apply it to real physical systems (as opposed to just the idealized "Turing Machine"). Nevertheless, there are enough people around (and I may be one of them) who also subscribe to the generalization of the C-TT to physical systems (to a close enough discrete approximation). That's the version of C-TT I had in mind, and the Computationalist Thesis that Cognition IS Computation (C=C) still does not follow from it.

It is of course possible that the extended C-TT is false, and hence (a fortiori, as you say) that the brain (and any other physical system capable of cognition) is NOT describable by computation. I don't happen to think this is the case, but it's certainly a possibility. The important logical point here, though, is still that C=C does NOT follow even from the extended C-TT (even though some people seem to think it does) and hence that one can (like Searle and me) subscribe to the extended C-TT while still rejecting C=C.

To put it yet another way, Godel-incompleteness arguments against C=C -- as invoked by, say, Lucas or Penrose -- are actually arguments against the extended C-TT for the special case of the human mind and brain, but Searlean arguments against C=C are not. Nor is the Symbol Grounding Problem a challenge to the extended C-TT; just to the conflation of the real/virtual distinction in the case of virtual minds -- symbol systems that are systematically interpretable as if they were thinking (Hayes et al. 1992).

>jd> (BTW, do you really mean describable _by_ computation rather than
>jd> _as_ computation?)

That distinction is too subtle for me. Let me state it more explicitly, using an analogy. Instead of computations (systems of formal symbols and symbol manipulations) let's speak of natural language. From the fact that a phenomemon is describable (to as close an approximation as you like) BY a set of sentences it does not follow that the phenomenon IS a set of sentences. Ditto for computational as opposed to linguistic descriptions. (By the way, I think computation is a subset of natural language, so this too is true "a fortiori"...) So thinking may be describable by/as (just implemented formal) computation, but it is not (just implemented formal) computation, any more than it is a set of sentences.

>
\>sh> So, to summarize, computational DESCRIBABILITY is not at issue;
>
\>sh> essential computationality is. To put it another way, what's at issue
>
\>sh> is the COMPUTATONALIST Thesis (that C = C), NOT the Church-Turing
>
\>sh> Thesis that (almost) everything is computationally describable and
>
\>sh> computationally equivalent to a UTM simulation. It's a mistake to
>
\>sh> conflate the two theses.
>
\>jd> Why do you think the Church-Turing Thesis says almost everything
\>jd> is computationally describable? All the Thesis says is that
\>jd> effectively computable = general recursive (or lambda-definable
\>jd> or computable by a Turing machine). -- Jeff Dalton

Agreed. It's the extended version of the C-TT I was referring to. I should not have written "anyone who subscribes to some version" but "anyone who subscribes to the extended version."

Stevan Harnad

---

The following article is retrievable by anonymous ftp from directory pub/harnad/Harnad on host princeton.edu

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on the Virtual Mind. Minds and Machines 2(3) 217-238. FILENAME: harnad92.virtualmind ABSTRACT: When certain formal symbol systems (e.g., computer programs) are implemented as dynamic physical symbol systems (e.g., when they are run on a computer), their activity can be interpreted at higher levels (e.g., binary code can be interpreted as LISP, LISP code can be interpreted as English, and English can be interpreted as a meaningful conversation). These higher levels of interpretability are called "virtual" systems. If such a virtual system is interpretable as if it had a mind, is such a "virtual mind" real? This is the question addressed in this "virtual" symposium, originally conducted electronically among four cognitive scientists: Donald Perlis, a computer scientist, argues that according to the computationalist thesis, virtual minds are real and hence Searle's Chinese Room Argument fails, because if Searle memorized and executed a program that could pass the Turing Test in Chinese he would have a second, virtual, Chinese-understanding mind of which he was unaware (as in multiple personality). Stevan Harnad, a psychologist, argues that Searle's Argument is valid, virtual minds are just hermeneutic overinterpretations, and symbols must be grounded in the real world of objects, not just the virtual world of interpretations. Computer scientist Patrick Hayes argues that Searle's Argument fails, but because Searle does not really implement the program: A real implementation must not be homuncular but mindless and mechanical, like a computer. Only then can it give rise to a mind at the virtual level. Philosopher Ned Block suggests that there is no reason a mindful implementation would not be a real one.

Date: Fri, 9 Jul 93 15:57:19 EDT From: "Stevan Harnad"

bc> From: bclarke@cogsci.UCSD.EDU (Bob Clarke) bc> Date: Fri, 9 Jul 93 8:14:40 PDT bc> bc> In reply to Harnad, I quote Chris Langton: bc> bc> "Is a school of fish a simulation of a flock of birds?"

Unfortunately, this misses the point. The question is not about simulation, description or analogy. The question is: IS Cognition Computation? The answer is "No," just as the answer to IS Swimming Flying? or IS Flying Swimming? would be "No."

Change the question and make it "Are there interesting similarities between swimming and flying, or between birds and fish?" and the answer may be "Yes." But Computationalism is not asking that question. It is asking whether cognition IS computation. Similarity is not identity (any more than computational equivalence is).

Stevan Harnad

Harnad, S. (1993) Artificial Life: Synthetic Versus Virtual. Artificial Life III. Proceedings, Santa Fe Institute Studies in the Sciences of Complexity. Volume XVI.

Date: Sun, 11 Jul 93 14:10:46 EDT From: "Stevan Harnad" Subject: Re: "Is Cognition Computation?"

rk> From: Robert.Kentridge@durham.ac.uk rk> Date: Sun, 11 Jul 1993 13:08:13 +0100 (BST) rk> To: harnad@Princeton.EDU (Stevan Harnad) rk> rk> As is often the case I don't think [Harnad and I] disagree much at all. rk> In my previous message I hoped to point out that a physical system rk> could have a potentially limitless number of valid computational rk> descriptions. I then noted, a bit flippantly, that understanding how rk> information was transduced between the outside world and a supposedly rk> cognizing system might inform us as to which of these computational rk> descriptions had something to do with cognitive function. I don't rk> think any of this commits me to believing that cognition is rk> computation. I also hadn't intended to imply that the only reason for rk> being interested in transduction is identifying functionally rk> significant computational descriptions of cognizing systems. As I rk> said in my Minds and Machines contribution I don't think we're dealing rk> with atomic value-free 'symbols' when we consider symbolic rk> descriptions of systems whose behaviour evolves via appropriately rk> transduced interaction with a real world.

rk> I'd be interested to know if we disagree over the following, rk> which is the implicit position which has underlain my messages in this rk> discussion. Cognition isn't just computation, but cognizing systems rk> can be described computationally. The nature of this computational rk> description of a system might be a necessary but not sufficient rk> condition for cognition to be occurring in that system. We can still rk> discover interesting things about processes and representations in rk> cognition by studying the computation going on in cognizing systems rk> (not the computation we think ought to be going on in them!) without rk> believing that cognition is just computation. rk> rk> Dr. R.W. Kentridge Psychology Dept. University of Durham,

Alas, I do agree with this completely (so there won't be any new new disagreement-induced breakthroughs here!), but Bob's position is exactly what Searle would have called "Weak AI," and it is uncontested. Note also that the exact same thing can be said of most of physics and engineering. This just means that cognitive theory, like most of the rest of scientific theory, is COMPUTABLE, not that thinking (or movement, or heat) are COMPUTATIONAL. That sounds like R.I.P for C=C.

Stevan Harnad

----

bc> Date: Fri, 9 Jul 93 23:13:30 -0700 bc> From: uunet!questrel.questrel.COM!chris@UCSD.EDU (Chris Cole) bc> Sender: bclarke@cogsci.UCSD.EDU bc> bc> I don't think Langton missed the point. If we agree that a flock of birds is bc> a simulation of a school of fish, and vice versa, then it seems we agree bc> that a flock of birds is a simulation of itself. The use-mention levels bc> Harnad wishes to observe vanish in the case of cognition.

Chris Langton and I disagree about synthetic vs. virtual life (as discussed in the Alife III article I cited earlier) but the above point doesn't sound like Langton's, nor does it sound entirely coherent.

If A is like B in some respect, then A is like B in some respect. There is some property, concrete or abstract, that they share. A fortiori, A is also like A in that respect. It too has that shared property. So what?

"Simulation" has two meanings. It can be used loosely to just mean "is like," the shared-property sense noted above-- in which case NOTHING follows from the above observations -- or it can be used in the technical sense of computer simulation, or computational equivalence. In that special sense, A might be a real state, like flying, and B might be a symbolic state in a computer, like virtual flying. That's where the disagreement between me and Chris resides, and it's not resolved by invoking shared properties of real flying and real swimming.

The issue there is real vs. virtual, and the question is about IDENTITY, not SIMILARITY. In this forum what is under discussion is whether cognition IS computation, not whether it is LIKE computation. In the Artificial-Life context, the disagreement is over whether virtual life IS alive, not whether it is LIKE things that are alive.

If you post further on this question, please first identify who you are (it's not clear whether you are B. Clarke or C. Cole) and indicate whether you have been following this discussion. This is not a bulletin board on which anything is gained by interjecting a comment after having read only one or a few exchanges in a thread. If you have not been following the discussion, please read the archive before sending further comments.

Stevan Harnad

Harnad, S. (1993) Artificial Life: Synthetic Versus Virtual. Artificial Life III. Proceedings, Santa Fe Institute Studies in the Sciences of Complexity. Volume XVI.

From: Selmer Bringsjord Date: Mon, 16 Aug 93 12:12:04 -0400

>
>sh> The point of Turing's party game was partly to eliminate BIAS
>
>sh> based on appearance. We would certainly be prepared to believe
>
>sh> that other organisms, including extraterrestrial ones, might have
>
>sh> minds, and we have no basis for legislating in advance what their
>
>sh> exteriors are or are not allowed to LOOK like (any more than we
>
>sh> can legislate in advance what their interiors are supposed to look
>
>sh> like). It's what they can DO that guides our judgment, and it's the
>
>sh> same with you and me (and the Blind Watchmaker: He can't read
>
>sh> minds either, only adaptive performance).

You here happily begin to beg the question against the *modus operandi* I represent. One can't simply *assume* that it's only what aliens can *do* that guides our judgement, for we may be able to sit down in our armchairs and reason things out well in advance of their arrival -- if we have sufficient information. If we hear in advance, for example, that the Nebulon race is composed of beings who are physical implementations of finite state automata, we can immediately deduce quite a bit about the cognitive capacity of Nebulons. Here's the same error rearing its head again:

>
>sh> ...if you could correspond with me as a lifelong pen-pal I have
>
>sh> never seen, in such a way that it would never cross my mind that
>
>sh> you had no mind, then it would be entirely arbitrary of me to
>
>sh> revise my judgment just because I was told you were a machine -
>
>sh> - for the simple reason that I know neither what PEOPLE are nor
>
>sh> what MACHINES are.

(By the way, if it never crosses S's mind that ~p, then S will in fact never revise her judgement that p, right?) Look, if I'm told that your pen-pal, Pen, is a (computing?) machine, *and* I reflect on whether or not Pen is a person, I'll quickly come to the conclusion that Pen isn't. Why? Because in order for Pen to be a person there must be something it's like to be her on the inside. But there's nothing it's like to be a computing machine on the inside. Hence Pen isn't a person. (Spelled out in Chapter I, of *What Robots Can & Can't Be*.) Because in order for Pen to be a person she must have the capacity to, within certain parameters, introspect infallibly. But no computing machine can have such a capacity. Hence Pen isn’t a person. (Spelled out in Chapter IX, of *WRC&CB*.) Because in order for Pen to be a person she must have free will. But no computing machine can have free will. Hence Pen isn’t a person. (Spelled out in Chapter VIII, of *WRC&CB*.) Etc., etc.

Now tell me what premises are false in these arguments, or tell me to flesh out the arguments on the spot, in plain English (which can’t be done, at least not well: more about that later), but don’t tell me that which convicts you of *petitio principii*: don’t tell me that one simply can’t deduce things about Pen from my Chesterfield. Whether or not one can is precisely what’s at issue! Do you see the vast methodological chasm that separates us here? You have even managed to box Searle’s CRA into your way of looking at the world of ideas, for you take CRA’s moral to be limited to its conclusion, while philosophers who affirm CRA see also a broader moral: sit down and think, calmly, carefully, without manipulating test tubes, without programming, just sit down and reason, and you can leapfrog and enlighten an empiricist crowd. This is why CRA just looks to me like one in a long line of arguments which are serving to turn "person-building" AI into a carcass, leaving alive only those engineers who will manage a living on the basis of systems which do helpful but (in the grand scheme of things) shallow things like direct air traffic read echocardiograms, and, in my own case, generate stories.

>
\>sh> Regarding the serendipity argument, let me quote from my reply
>
\>sh> to your commentary (Bringsjord 1993) on Harnad (1993a): "So
>
\>sh> chimpanzees might write Shakespeare by chance: What light does
>
\>sh> that cast on how Shakespeare wrote Shakespeare?" This is
>
\>sh> perhaps the difference between an engineering and a
>
\>sh> philosophical motivation on this topic. Physicists don’t worry
>
\>sh> about thermodynamics reversing by chance either, even though
>
\>sh> it’s a logical possibility. It’s just that nothing substantive rides
>
\>sh> on that possibility. A life-long pen-pal correspondence could be
>
\>sh> anticipated by chance, logically speaking. So what? What we are
>
\>sh> looking for is a mechanism that does it by design, not by chance.

You’re targeting a straw man. I don’t maintain anything even remotely like the claim that the possibility that chimpanzees write Shakespeare by chance casts light on how Shakespeare wrote Shakespeare! Who in their right mind would ever promote this silly idea? What I’ve said is that thought-experiments exploiting serendipitous passing of TT show that certain construals of Turing’s main thesis are false, because these construals put the thesis in the form of a conditional, and the thought-experiments are cases in which the antecedent is true but the consequent isn’t.

There *is* a difference in motivation on the topic -- you're right here. But I've tried to explain that I have *both* motivations, both the philosophical and the engineering. As philosopher I'm concerned with truth (the truth of Turing's, or Harnad's, thesis concerning TTs). As engineer I'm concerned with building a system that passes at least some of the TTs. I'm quite confident (as you know, from reading WRC&CB) that TT-passing robots will arrive on the scene. Whether these robots are persons is another matter.

So, when you say

>
>sh> A life-long pen-pal correspondence could be anticipated by
>
>sh> chance, logically speaking. So what? What we are looking for is
>
>sh> a mechanism that does it by design, not by chance.

I'm afraid you conflate the philosophical w/ the engineering. The engineer is of course looking for a mechanism that does it by design. The philosopher is looking for a truth-value for a mechanism that does it that is conscious.

Now this thing about conditionals really needs to be cleared up. While you say

>
>sh> The only conditional I've found useful here so far is Searle's: "IF a
>
>sh> pure symbol-cruncher could pass T2, it would not understand,
>
>sh> because I could implement the same symbol-cruncher without
>
>sh> understanding."

here's a conditional you're apparently proposing:

(1.0) If x pen-pals w/ Harnad for 20 yrs, then x is conscious.

Look at some of the variations which arise before we even think about turning for help to formalisms:

(1.1) If x pen-pals w/ Harnad for 20 yrs, then x is probably conscious.

(1.2) If x pen-pals w/ Harnad for 20 yrs, then Harnad wouldn't be irrational in holding that x is conscious.

(1.3) If x pen-pals w/ Harnad for 20 yrs, then Harnad would hold that x is conscious.

(1.4) It's physically necessary that (1.i).

(1.5) It's logically necessary that (1.i).

etc., etc.

What is it that you mean to champion w.r.t. to pen-pals and TTT? At one point I thought for sure that you meant to push

(2) If a robot *r* passes the Total Turing Test, then we are no more (or less) justified in denying that *r* is conscious then we are justified in denying that *h*, a human, observed in ordinary circumstances, is conscious.

I believe I showed why this proposition is false in my *Think* commentary. So I don't really know what your position is. I know what you want to build -- we want to build the same thing. But can you simply express in one declarative sentence what your position on consciousness and Turing testing is? We can then analyze that sentence with help from the customary sources and try to see if it's true.

I realize that you want to promote a situation wherein we sort of ignore truth-values. That's why you say

>
>sh> I can only repeat, "For every x, if x passes TT, then x is conscious"
>
>sh> was just a conjecture, with some supporting arguments. It turns
>
>sh> out to be very probably false (unless memorizing symbols can
>
>sh> make Searle understand Chinese, or generate a second mind in
>
>sh> him that understands Chinese). No Searlean argument can be
>
>sh> made against T3, but that could of course be false too; and so
>
>sh> could even T4. So forget about proofs or necessity here. It's
>
>sh> underdetermination squared all the way through, because of
>
>sh> what's abidingly special about the mind/body problem (or about
>
>sh> qualia, if you prefer).

But this is rather self-serving, it seems to me. You can't legislate for yourself a position which can't be scrutinized in the customary way by philosophers. Philosophers don't leave propositions like "God exists" alone as conjectures; they ain't gonna leave your position on Turing testing and consciousness alone either -- not for a second. You put all the weight, it seems, on a robot's sensors and effectors. Does that somehow insulate you from arguments purporting to show that featuring sensors and effectors doesn't help? To repeat, if Turing's conditional can be counter-exampled in the standard way (by finding a scenario wherein the antecedent is true but the consequent is false -- that's what my argument from serendipity involves), I see no reason why your conditional can't be done in similarly. Give me the conditional and let's see. But I should warn you, you give every indication that in your pocket resides a bankrupt conditional.

Of course, you also indicate that though you may affirm the conditional, you somehow don't do it publicly. That's how I understand the somewhat mysterious

>
\>sh> ...I never endorsed the positive version of T2 or T3! For me, they
>
\>sh> are epistemic, not ontic criteria (empirical constraints on models,
>
\>sh> actually).

That is, I hear you saying here that you affirm some relevant (1.i)- ish conditional on Total Turing Testing, but you don't really want to come right out and admit it. Is this where you stand?

>
\>sh> Chapter IX is too much of a thicket. The CRA is perspicuous; you
>
\>sh> can say what's right or wrong with it in English, without having to
>
\>sh> resort to formalisms or contrived and far-fetched sci-fi scenarios.
>
\>sh> I have restated the point and why and how it's valid repeatedly
>
\>sh> in a few words; I really think you ought to do the same.

Hmm. I think you've got things reversed. Chapter IX's (actually: Ch. VIII's) argument from free will is a formally valid argument -- that much can be proved. It's conclusion is that person-building AI is doomed. Since it's formally valid, if its premises are true, person- building AI is doomed. Your response to this is that it's too much of a thicket? When people say that about Searle's CRA, they need to be walked through a very carefully stated version of the argument -- a version Searle never produced. You simply buy Searle's *idea*. You look at Searle's thought-experiment and you say, Yes! That's fortunate for Searle, but rather uncommon. Searle has no qualms using terms like "syntax" and "semantics" baldly in his argumentation. "Semantics," if left all alone, is on par with "free will" or "God." Chapter VIII takes free will, and works it out. The basic idea can be expressed in Searlean fashion rather easily, but, in the absence of a formal follow-up, it seems worthless to. At any rate, here 'tis:

HIGH-LEVEL ENCAPSULATION OF CHAPTER VIII, *WHAT ROBOTS CAN & CAN'T BE*

(1) If every event is causally necessitated by prior events, then no one ever has power over any events (because, e.g., my eating pizza was ineluctably in the cards when T. Rex roamed the planet).

(2) If it's not the case that every event is causally necessitated by prior events, then, unless *people, not events, directly bring about certain events, and bring about their bringing about those events* (the doctrine is known as iterative agent causation), no one ever has power over any state of affairs (because, e.g., my eating pizza would then simply happen uncaused and "out of the blue," and would thereby not be anything over which I have power).

(3) Either every event is causally necessitated by prior events, or not (a tautology!).

Therefore:

(4) Unless iterative agent causation is true, no one ever has power over any events.

(5) If no one ever has power over any events, then no one is ever morally responsible for anything that happens.

(6) Someone is morally responsible for something that happens.

Therefore:

(7) It’s not the case that no one ever has power over any events.

Therefore:

(8) Iterative agent causation is true.

(9) If iterative agent causation is true, then people can’t be automata (because if people can enter an infinite amount of states in a finite time, were they automata, they would be able to solve unsolvable computational problems).

Therefore:

(10) People aren’t automata.

>
\>sh> (Also, speaking as [as far as I know] the first formulator of the T-
>
\>sh> hierarchy, there is, despite your hopeful dots after T2, T3..., only
>
\>sh> one more T, and that’s T4! The Turing hierarchy (for mind-
>
\>sh> modelling purposes) ENDS there, whereas the validity of CRA
>
\>sh> begins and ends at T2.)

The T1-T4 hierarchy (could you refresh me on that? -- we used to call it TT, TTT, TTTT) is too restrictive, mathematically speaking. That’s why Kugel has considered tests which no finite state automaton can pass (a finite state automaton can pass T1-T4; and Searle operates as an FSA in the CRA). I’ve discussed the Turing Test Sequence under the assumption that after T1-T4 can come any number of tests designed to get at different grades of COMPUTATIONAL power.

>
\>sh> This has nothing to do with nonconstructive vs. constructive proof.
>
\>sh> As I understand it, CRA is not and cannot be a proof. If you can
>
\>sh> upgrade it to one, say how, in a few transparent words. Students

>
>sh> can be persuaded in many ways; that's irrelevant. I've put my
>
>sh> construal briefly and transparently; you should do the same.

Stevan. On the contrary, there is a certain clash in the foundations of mathematics which casts considerable doubt on the definition of PROOF you're evidently presupposing (because it shows that proof is a person- and group-varying notion). You don't understand CRA to be a proof, but I claim that when CRA is modified slightly and formulated more carefully it *is* a proof. My definition of proof is a formally valid chain of reasoning P1, P2, P3, ..., Pn such that a significant number of those in community involved affirm Pn on the strength of P1-Pn-1. Let me give you a quote that may prove somewhat illuminating here:

"Stanislaw Ulam estimates that mathematicians publish 200,000 theorems every year. A number of these are subsequently contradicted or otherwise disallowed, others are thrown into doubt, and most are ignored. Only a tiny fraction come to be understood and believed by any sizable group of mathematicians. The theorems that get ignored or discredited are seldom the work of crackpots or incompetents. In 1879, Kempe published a proof of the four-color conjecture that stood for eleven years before Heawood uncovered a fatal flaw in the reasoning..." (many other examples follow). [DeMilo, Lipton & Perlis, RSocial Processes and Proofs of Theorems and Programs,S CACM 22, 1979]

This means that "proofs" are regularly published which contain reasoning that isn't formally valid. In *What Robots Can & Can't Be* I have produced arguments most of which are formally valid, and, it seems to me, most of which have very plausible premises. The only missing ingredient is affirmation by others. Over that, I have no control. But notice that a "proof" is affirmed as part of what makes it a proof. That's why I regard student reaction to be significant. Most of these students have fairly open minds: they're not heavily invested in AI dogma. The cognition is computation thesis, put explicitly, is for them a fresh claim.

Your feeling that arguments can be upgraded in a few words to a proof is peculiar. If anything, it's the other way around: proofs can be downgraded to mere sketches when they are expressed in a few words. Godel's incompleteness theorems are a perfect case in point: to grasp the theorems themselves is necessarily to grasp every point along the *actual* proof. I get students all the time who think they understand Godel's results solely on the strength of popularizations. None of them do, and they come to realize this when they roll up their sleeves and get into the details.

We fight a continuing methodological battle along these lines, of course. You live for the supple, elegant, clear statement of an argument in English. I don't think such a medium gives anything but a feint echo of the truth, and so live for the painstakingly explicit statement of an argument in at least an English-logic hybrid :-).

CHRISLEY-HARNAD EXCHANGE

This exchange revolves around three theses: (CTT), (CTT-P), (C=C), the Church-Turing Thesis, the "physics version" of (CTT), and "cognition is computation," resp. As such, the exchange revolves around massive falsehood. In "Church's Thesis, *Contra* Mendelson, is Unprovable, and Worse ... It May be False," I show how I'll soon counter-example (CTT). This paper will be presented at the annual upcoming Eastern Division meeting of the American Philosophical

Association in Atlanta (all are invited). (CTT), it seems to me, as well as to more than a few physicists I know, *can't* be applied to the world of the quantum (ergo, (CTT-P) falls). As for (C=C), I take my *What Robots Can & Can't Be* to be a refutation. Though I may be accused of kicking a dead horse, Michael Zenzen and I are now refining a brand new sort of attack on (C=C) -- no Searle, no Nagelian qualia, something brand new. Along the way we deploy something you have elegantly expressed, Stevan, when you say to Chrisley: "...the whole point of computationalism was to partition phenomena into those that are fully described only by their physics, and those that are fully described by their *computations*, independent of their physics."

Selmer Bringsjord

From harnad Sat Aug 21 18:16:40 1993 To: Symbol Grounding Discussion Group:

TURING'S TEST IS A METHODOLOGICAL CONSTRAINT AND SEARLE'S CRITIQUE IS A PLAUSIBILITY ARGUMENT: NEITHER IS AMENABLE TO PROOF

Selmer Bringsjord wrote (regarding Turing Testing):

>sb> One can't simply *assume* that it's only what aliens can
>sb> *do* that guides our judgement, for we may be able to sit down in our
>sb> armchairs and reason things out well in advance of their arrival...
>sb> If we hear in advance, for example, that the Nebulon race is composed
>sb> of beings who are physical implementations of finite state automata, we
>sb> can immediately deduce quite a bit about the cognitive capacity of
>sb> Nebulons.

It MIGHT be possible in principle to prove that certain kinds of T3-passers (Turing-indistinguishable from us in their robotic and symbolic performance capacities) can't have minds, but if so, I haven't a clue how such a proof might run. For, even in the obvious cases (cheating by remote human telemetric control of the robot, or a robot that happens to do it all purely by chance), the cases one would want to reject a priori, the logical POSSIBILITY that the candidate has a mind anyway cannot be eliminated. The same is true of Searle's argument that a T2-passing computer would not have a mind: No proof, it only suggests that it's extremely unlikely. I don't see why you want to turn these epistemic and methodological questions into matters for PROOF, rather than mere arguments for plausibility on the evidence. It seems to me that the other-minds barrier (the impossibility of knowing or showing with certainty that a system has [or hasn't] a mind other than by being the system) effectively rules out proofs (it's already a monkey-wrench even for ordinary empirical inference).

I wish other readers would jump in here, but, on my reading, what you keep doing is formalizing very simple statements until one (or I, at least) can no longer keep track of the mnemonics of what each means, and then you draw formal conclusions that I would certainly not endorse from my original construal of the statements. This would be fine if the subject matter here were MORE complicated, and the notation were to SIMPLIFY and CLARIFY it, but what I find instead here is that all the complication comes from the notation itself, and at some point the intended interpretation of the original statements is lost, and we are simply following lockstep with some algebra that no longer has any connection with what we meant.

Let me put it simply here with reference to what you said above about what might be provable a priori about Nebulons and performance capacity: I cannot imagine a formal argument you could use for the fact that if a system passed T3 but were of a particular kind (say, a finite state automaton) then it COULD NOT (with the force of either logical or physical necessity) have a mind -- for the simple reason that (for all I know, or can even conceive of knowing) even a rock or an electron could have a mind.

I could, on the other hand, easily imagine a proof that a finite state automaton could not pass T3 -- but that's an entirely different story, a performance engineering story. There's a difference like night and day between trying to prove that something does or not have certain observable properties and trying to "prove" that something does or does not have a mind.

>sb> Look, if I'm told that your pen-pal, Pen, is a (computing?) machine,
>sb> *and* I reflect on whether or not Pen is a person, I'll quickly come to
>sb> the conclusion that Pen isn't. Why? Because in order for Pen to be a
>sb> person there must be something it's like to be her on the inside. But
>sb> there's nothing it's like to be a computing machine on the inside.
>sb> Hence Pen isn't a person. (Spelled out in Chapter I, of *What Robots
>sb> Can & Can't Be*.) Because in order for Pen to be a person she must have
>sb> the capacity to, within certain parameters, introspect infallibly. But
>sb> no computing machine can have such a capacity. Hence Pen isn't a
>sb> person. (Spelled out in Chapter IX, of *WRC&CB*.) Because in order for
>sb> Pen to be a person she must have free will. But no computing machine
>sb> can have free will. Hence Pen isn't a person. (Spelled out in Chapter
>sb> VIII, of *WRC&CB*.) Etc., etc.

Alas, I have read all these chapters, but I do NOT find them to support any of the conclusions you adduce above.

>sb> Now tell me what premises are false in these arguments, or tell me to
>sb> flesh out the arguments on the spot, in plain English (which can't be
>sb> done, at least not well: more about that later), but don't tell me that
>sb> which convicts you of *petitio principii*: don't tell me that one
>sb> simply can't deduce things about Pen from my Chesterfield. Whether or
>sb> not one can is precisely what's at issue!

Alas, I have to say exactly what you enjoin me not to: For a starter, how is any formal argument going to penetrate the other-mind barrier with the force of proof? (I've already agreed that cheating, chance and the Chinese Room Argument against a T2-passing computer have the force of a plausibility argument, but PROOF?)

>sb> What I've said is that thought-experiments exploiting serendipitous
>sb> passing of TT show that certain construals of Turing's main thesis are
>sb> false, because these construals put the thesis in the form of a
>sb> conditional, and the thought-experiments are cases in which the
>sb> antecedent is true but the consequent isn't.

But I immediately concede that the conditional that "if X passes T2 (or T3) it NECESSARILY has a mind" is false, because the passing could be a result of chance, cheating, or mindless symbol-manipulation (Searle). The first two of those counterexamples are less an invalidation of T2/T3 as evidence than of T2/T3 as proof. (Searle's argument also invalidates T2 as evidence, in the special case of a [hypothetical] implementation-independent T2-passing symbol system). No formalism was needed to see any of this.

I must also repeat that my own construal of T3 is as a methodological criterion, not a proof, a criterion that is best stated in the negative:

We (so far) know nothing about persons or machines that gives us any nonarbitrary reason for REVISING (if we should happen to learn that the candidate is a machine) our natural inference about the candidate whose (lifelong) performance capacity is T3-indistinguishable from that of a person with a mind.

This is no longer true for T2 in the special case just mentioned (implementation-independent symbol manipulation): there we DO have nonarbitrary reasons, namely, Searle's argument and the symbol grounding problem. These nonarbitrary reasons are transparent to us all (though, being mere claims about plausibility rather than proofs, they can still be rejected -- and indeed they ARE still rejected by "computationalists" -- but at the price of a lot of sci-fi special pleading, in my opinion).

What I fail to get from your formal "proofs" is the counterpart of the transparent reasons on which Searle's argument and my formulation of the symbol grounding problem draw. If I could see some of those, perhaps I could better appreciate the thrust of the proofs. But for now, they just seem to be formal symbol manipulations! The intended interpretation of your symbols has somehow fallen by the wayside, or gotten lost in the thicket of mnemonics and their variants.

>sb> here's a conditional you're apparently proposing:
>sb> (1.0) If x pen-pals w/ Harnad for 20 yrs, then x is conscious.

Not at all. I say only that "If x pen-pals w/ Harnad for 20 yrs AND x is just the (irrelevant) implementation of a symbol system (every other implementation of which is also conscious if x is conscious), then x is not conscious."

The (so far) unassailed T3 version is: "If x walketh T3-indistinguishably among us for a lifetime, then, if informed that x is a machine, we have no nonarbitrary basis for denying of x what we affirm of one another, namely, that x is conscious." Of course, x might still fail to be conscious, for example, if, unbeknownst to us, x was controlled telemetrically by someone else, who WAS conscious, or if, mirabile dictu, x performed T3-indistinguishably by chance. But that just about exhausts potential nonarbitrary reasons for denying that x was conscious -- though it does not exhaust all possible (hidden) CAUSES for x's failing to be conscious. I doubt that we can ever KNOW those hidden causes in the usual empirical way, however; for even if science shows that organisms always pass T3 with one kind of internal mechanism, and only synthetic robots pass it with another kind (this difference is a difference in T4), we could never be sure the robots weren't conscious too (and I doubt you could "prove" they weren't).

Turing Indistinguishability scales up seamlessly to Empirical Indistinguishability (Harnad 1992). This is where the metaphysics of indiscernibility comes up against the special case of the first-person point of view...

>sb> Look at some of the variations which arise before we even think
>sb> about turning for help to formalisms:
>sb> (1.1) If x pen-pals w/ Harnad for 20 yrs, then x is probably
>sb> conscious.
>sb> (1.2) If x pen-pals w/ Harnad for 20 yrs, then Harnad wouldn't be
>sb> irrational in holding that x is conscious.
>sb> (1.3) If x pen-pals w/ Harnad for 20 yrs, then Harnad would hold
>sb> that x is conscious.
>sb> (1.4) It's physically necessary that (1.i).
>sb> (1.5) It's logically necessary that (1.i). etc., etc.
>sb> What is it that you mean to champion w.r.t. to pen-pals and TTT?

I've spelled what I champion above; none of these variants are relevant or necessary; nor do they need names (or numbers, or acronyms).

>sb> At one point I thought for sure that you meant to push:
>sb> (2) If a robot *r* passes the Total Turing Test, then we are no more
>sb> (or less) justified in denying that *r* is conscious then we are
>sb> justified in denying that *h*, a human, observed in ordinary
>sb> circumstances, is conscious.
>sb> I believe I showed why this proposition is false in my *Think*
>sb> commentary. So I don't really know what your position is. I know
>sb> what you want to build -- we want to build the same thing. But can
>sb> you simply express in one declarative sentence what your position
>sb> on consciousness and Turing testing is? We can then analyze that
>sb> sentence with help from the customary sources and try to see if it's
>sb> true.

Your statement above is close enough. Needless to say, I don't think you've shown it to be false in your commentary (which is ftp-retrievable from princeton.edu as the file pub/harnad/Harnad/harnad93.symb.anal.net.bringsjord). I would immediately concede that if we know it passed T3 by "serendipity" or trickery then that WOULD be justification for denial, but so what? I am not talking about proof but about reasons and plausibility.

>sb> To repeat, if Turing's conditional can be counter-exampled in the
>sb> standard way (by finding a scenario wherein the antecedent is true but
>sb> the consequent is false -- that's what my argument from serendipity
>sb> involves), I see no reason why your conditional can't be done in
>sb> similarly.

Done. So what? It never had the force of proof. And all it called for was a nonarbitrary reason. Besides, cheating or chance could have done in the case for F = ma (or quarks) too (not that I'm implying that the case for T3 and consciousness is merely like that: I think the special features of the mind/body problem make the case of qualia very different from that of quarks, even though both are unobservable; see Harnad 1993). This just isn't the kind of question that there is any need or justification for resorting to formalism in order to address. And it’s not a fruitful terrain for formal "proofs."

>
>sh> ...I never endorsed the positive version of T2 or T3! For me, they
>
>sh> are epistemic, not ontic criteria (empirical constraints on models,
>
>sh> actually).

>sb> That is, I hear you saying here that you affirm some relevant (1.i)-
>sb> ish conditional on Total Turing Testing, but you don’t really want to
>sb> come right out and admit it. Is this where you stand?

No. The positive version is:

"X has a mind if it is T3-indistinguishable from someone with a mind."

I think that’s unjustified. The negative version I do endorse is:

"If X is T3-indistinguishable from someone with a mind, one has no nonarbitrary reason for doubting that X has a mind when told that X is a machine."

Except if one suspects chance or cheating, of course. (These are ordinary ceteris paribus conditions that I think no roboticist -- indeed no scientist -- needs to mention explicitly.)

>sb> I think you’ve got things reversed. [My] Chapter... VIII’s argument
>sb> from free will is a formally valid argument -- that much can be proved.
>sb> Its conclusion is that person-building AI is doomed. Since it’s
>sb> formally valid, if its premises are true, person-building AI is
>sb> doomed. Your response to this is that it’s too much of a thicket? When
>sb> people say that about Searle’s CRA, they need to be walked through a
>sb> very carefully stated version of the argument -- a version Searle never
>sb> produced. You simply buy Searle’s *idea*. You look at Searle’s
>sb> thought-experiment and you say, Yes! That’s fortunate for Searle, but
>sb> rather uncommon...

Searle’s argument is transparent. In stating it more carefully than Searle did (as I have tried to do), one makes it even more transparent. Your version, on the other hand, makes it more opaque (if it’s even the same argument at all -- I can’t tell.)

>sb> HIGH-LEVEL ENCAPSULATION OF CHAPTER VIII, *WHAT ROBOTS CAN & CAN’T BE*:
>sb> (1) If every event is causally necessitated by prior events,
>sb> then no one ever has power over any events (because,
>sb> e.g., my eating pizza was ineluctably in the cards when T.
>sb> Rex roamed the planet).

(I happen to think the antecedent here, and hence the consequent, is true, but I certainly don’t see any reason to make anything in robot-building depend on any such controversial metaphysical conjectures.)

>sb> (2) If it’s not the case that every event is causally
>sb> necessitated by prior events, then, unless *people, not
>sb> events, directly bring about certain events, and bring
>sb> about their bringing about those events* (the doctrine is
>sb> known as iterative agent causation), no one ever has
>sb> power over any state of affairs (because, e.g., my eating
>sb> pizza would then simply happen uncaused and "out of the
>sb> blue," and would thereby not be anything over which I
>sb> have power).

This stuff is almost more controversial than the T-testing itself. How can one imagine grounding CONCLUSIONS about T-testing on such premises? "iterative agent causation," sui generis causality! I’m just interested in whether T3 is a reliable guide for mind-modelling...

>sb> (3) Either every event is causally necessitated by prior
>sb> events, or not (a tautology!).

Yes, but also, I think, a red herring here. ("You either have stopped beating your wife, or not" is a tautology too...)

>sb> Therefore:
>sb> (4) Unless iterative agent causation is true, no one ever has
>sb> power over any events.
>sb> (5) If no one ever has power over any events, then no one is
>sb> ever morally responsible for anything that happens.

The introduction of yet another conjecture, this time ethical instead of metaphysical, and of still more controversial concepts (power over events, moral responsibility)... This, together with the formalism, is what I mean by making things more opaque istead of clearer.

>sb> (6) Someone is morally responsible for something that happens.

Many are ready to contest this. But look, this "proof" is bringing in more and more controversial stuff.

>sb> Therefore:
>sb> (7) It’s not the case that no one ever has power over any events.
>sb> Therefore:
>sb> (8) Iterative agent causation is true.
>sb> (9) If iterative agent causation is true, then people can’t be
>sb> automata (because if people can enter an infinite amount
>sb> of states in a finite time, were they automata, they would
>sb> be able to solve unsolvable computational problems).

Doesn’t follow. They may not be able to enter the "right" states to solve the mathematical problems. And "iterative agent causation," whatever its pedigree, sounds pretty dubious to me.

>sb> Therefore:
>sb> (10) People aren’t automata.

All the force of this argument was in the various premises introduced. Unlike the axioms of arithmetic or geometry, these premises are FULL of things to disagree with, some of them even more controversial than the Turing Test itself. So how can you hope to base a "proof" of the validity or invalidity of the Turing Test on them?

And look: From Searle’s argument I have learned that people aren’t just implementation-independent implementations of formal symbol systems. Even if I really learned from your argument that people weren’t "automata," what would that make people instead? and what is one to do next? Searle’s argument and the symbol grounding problem turned us from the T2-symbolic road to the T3-hybrid road. What would be the corresponding guidance we would derive from this "proof" that people aren’t automata?

>sb> The T1-T4 hierarchy (could you refresh me on that? -- we used to
>sb> call it TT, TTT, TTTT) is too restrictive, mathematically speaking.
>sb> That’s why Kugel has considered tests which no finite state
>sb> automaton can pass (a finite state automaton can pass T1-T4; and
>sb> Searle operates as an FSA in the CRA). I’ve discussed the Turing Test
>sb> Sequence under the assumption that after T1-T4 can come any
>sb> number of tests designed to get at different grades of
>sb> COMPUTATIONAL power.

There is only T2 - T4. The "T" is for "Total" as in "Totally indistinguishable." There is the standard Turing Test (TT, T2): the successful candidate must be totally indistinguishable from us in its symbolic performance capacity. Then there is the "Total Turing Test" (TTT, T3): totally indistinguishable from us both symbolically and robotically. And last there is T4: totally indistinguishable from us symbolically, robotically, and neurally. After that, there’s nothing left! The T-hierarchy is just an EMPIRICAL hierarchy for reverse engineering (Harnad 1994). Because of the extra "T" for Turing, "T1" is empty. It’s best thought of as "t1" -- a toy model or module that is not Totally (or Turing-Indistinguishably) anything. It is Partial, underdetermined, hence potentially irrelevant. The very purpose of the Turing hierarchy is to lead us AWAY from arbitrary, underdetermined, subtotal, toy models.

So I don’t know what hierarchy you are imagining, but it does not seem to be an extrapolation of the same dimension that T2 - T4 are on, for T2 - T4 pick out three respects in which a candidate’s performance could be indistinguishable from ours, but after symbolic, robotic and neural indistinguishability there’s nothing left! There’s no point speaking of candidates that are Turing-Indistinguishable from us in performance capacities we DON’T have; and there’s even less point in talking about Turing indistinguishability from some formal computing capacity, since, insofar as we are concerned, inner computational power is hypothetical, whereas external performance power is factual. Turing-indistinguishability from some hypothetical computational power that is attributed to us completely loses the spirit of Turing’s intuition, which was about about indistinguishability of OBSERVABLES.

By the way, if an "FSA" can pass T3 (total symbolic + robotic capacity), it already exceeds the scope of Searle’s argument, which applies only to T2 (total symbolic [pen-pal] capacity) and implementations of implementation-independent symbol systems. And T4 (symbolic, robotic and neural indistinguishability) and brains correspond to the approach Searle ADVOCATES over T2 and physical symbol systems, the one he assails. I advocate instead T3 and hybrid symbolic/nonsymbolic systems. If your FSA covers all of these, then it is beside the point. Are sensorimotor transducers FSAs (they're not symbol systems)? Are brains?

>
>sh> As I understand it, CRA is not and cannot be a proof. If you can
>
>sh> upgrade it to one, say how, in a few transparent words. Students
>
>sh> can be persuaded in many ways; that's irrelevant. I've put my
>
>sh> construal briefly and transparently; you should do the same.

>sb> Stevan. On the contrary, there is a certain clash in the foundations of
>sb> mathematics which casts considerable doubt on the definition of PROOF
>sb> you're evidently presupposing (because it shows that proof is a person-
>sb> and group-varying notion). You don't understand CRA to be a proof, but
>sb> I claim that when CRA is modified slightly and formulated more
>sb> carefully it *is* a proof. My definition of proof is a formally valid
>sb> chain of reasoning P1, P2, P3, ..., Pn such that a significant number
>sb> of those in community involved affirm Pn on the strength of P1-Pn-1.

But the chain of reasoning you gave above (about free will, iterative causation, moral responsibility, etc.) is hardly one on which there is community affirmation.

>sb> This means that "proofs" are regularly published which contain
>sb> reasoning that isn't formally valid. In *What Robots Can & Can't Be* I
>sb> have produced arguments most of which are formally valid, and, it seems
>sb> to me, most of which have very plausible premises. The only missing
>sb> ingredient is affirmation by others. Over that, I have no control. But
>sb> notice that a "proof" is affirmed as part of what makes it a proof.
>sb> That's why I regard student reaction to be significant. Most of these
>sb> students have fairly open minds: they're not heavily invested in AI
>sb> dogma. The cognition is computation thesis, put explicitly, is for them
>sb> a fresh claim.

Accuse me of playing Tortoise to your Achilles if you like, but I find the proofs in your book, be they ever so valid on the premises, irrelevant to the issues at hand and not at all binding on the conclusions I have been inclined to draw. I find many of the premises themselves, and even the elements of which they are composed, likewise uncompelling.

>sb> Your feeling that arguments can be upgraded in a few words to a proof
>sb> is peculiar. If anything, it's the other way around: proofs can be
>sb> downgraded to mere sketches when they are expressed in a few words.
>sb> Godel's incompleteness theorems are a perfect case in point: to grasp
>sb> the theorems themselves is necessarily to grasp every point along the
>sb> *actual* proof. I get students all the time who think they understand
>sb> Godel's results solely on the strength of popularizations. None of

>sb> them do, and they come to realize this when they roll up their sleeves
>sb> and get into the details.

Yes, but there’s nothing like Goedel’s proof or any other substantive, technical proof in your book. I agree that mathematical proofs can be oversimplified and their essential insights and rigor lost when they are stated in discursive form. But the reverse is also true. Simple discursive insights can be lost when they are turned into obscure formalisms.

>sb> We fight a continuing methodological battle along these lines, of
>sb> course. You live for the supple, elegant, clear statement of an
>sb> argument in English. I don’t think such a medium gives anything but a
>sb> faint echo of the truth, and so live for the painstakingly explicit
>sb> statement of an argument in at least an English-logic hybrid :-).

I think formal mathematical and logical problems should be formulated and solved by formal means. The question of whether a candidate that is to various degrees empirically indistinguishable from ourselves has a mind is not a mathematical or logical problem (though how to generate its performance capacity might be). Hence I think that formalizing it obcsures rather than clarifies.

>sb> CHRISLEY-HARNAD EXCHANGE

>sb> This exchange revolves around three theses: (CTT), (CTT-P), (C=C),
>sb> the Church-Turing Thesis, the "physics version" of (CTT), and
>sb> "cognition is computation," resp. As such, the exchange revolves
>sb> around massive falsehood. In "Church’s Thesis, *Contra* Mendelson,
>sb> is Unprovable, and Worse ... It May be False," I show how I’ll soon
>sb> counter-example (CTT). This paper will be presented at the annual
>sb> upcoming Eastern Division meeting of the American Philosophical
>sb> Association in Atlanta (all are invited). (CTT), it seems to me, as well
>sb> as to more than a few physicists I know, *can’t* be applied to the
>sb> world of the quantum (ergo, (CTT-P) falls).

I happen to subscribe to the Church-Turing Thesis about computability. I think it probably also applies to all discrete (difference-equation-governed) physical systems. Probably classical Newtonian continuous systems (differential-equation-governed) already exceed its scope; quantum systems would then exceed it a fortiori. But I also think none of this is relevant to mind-modelling, or T-testing -- EXCEPT for its bearing on the question we seem to agree upon below:

>sb> As for (C=C), I take my
>sb> *What Robots Can & Can’t Be* to be a refutation. Though I may be
>sb> accused of kicking a dead horse, Michael Zenzen and I are now
>sb> refining a brand new sort of attack on (C=C) -- no Searle, no Nagelian
>sb> qualia, something brand new. Along the way we deploy something
>sb> you have elegantly expressed, Stevan, when you say to Chrisley:
>sb> "...the whole point of computationalism was to partition phenomena
>sb> into those that are fully described only by their physics, and those
>sb> that are fully described by their *computations*, independent of
>sb> their physics."

There will be a further posting from Ron Chrisley shortly. But I do agree with you that implementation-independence (i.e., physics-independence) is an ESSENTIAL feature both of formal computation and of the "computationalist" theory of mind. Any move away from implementation-independence is a move away from formal computation, the Church-Turing Thesis, and the computational theory of mind, even if Ron insists on calling this new hybrid thing "computation." One can't have one's cake and eat it too.

Stevan Harnad

Harnad, S. (1992) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. SIGART Bulletin 3(4) (October) 9 - 10.

Harnad S. (1993) Discussion (passim) In: Bock, G.R. & Marsh, J. (Eds.) Experimental and Theoretical Studies of Consciousness. CIBA Foundation Symposium 174. Chichester: Wiley

Harnad, S, (1994) Does the Mind Piggy-Back on Robotic and Symbolic Capacity? To appear in: H. Morowitz (ed.) "The Mind, the Brain, and Complex Adaptive Systems.

From: David Powers Date: Fri, 20 Aug 1993 11:03:40 +0200 (MET DST)

Selmer Bringsjord writes:

>sb> Because in order for Pen to be a person she must have the capacity to,
>sb> within certain parameters, introspect infallibly. But no computing
>sb> machine can have such a capacity. Hence Pen isn't a person. (Spelled
>sb> out in Chapter IX, of *WRC&CB*.) Because in order for Pen to be a
>sb> person she must have free will. But no computing machine can have
>sb> free will. Hence Pen isn't a person. (Spelled out in Chapter VIII, of
>sb> *WRC&CB*.) Etc., etc.
>sb>
>sb> Now tell me what premises are false in these arguments ...

Fine! On this basis, I am not a person as I cannot introspect infallibly within the cadre of any parameters. On the other hand it is trivial for a computer to have this capability, for example what a debugger does is rehearses infallibly the engendering of a particular behaviour, and it is infallible within certain parameters, in particular in the absence of process external influences (asynchrony, race conditions etc.)

Thus a computer is not a computing machine, but may be a person, and a human is not a person, but may be a computing machine!

Moreover, the Cognitive Science interest in computers comes largely from precisely the transparency with which the simulation of behaviour is carried out by a computer, although there may indeed be parameters beyond which this transparency disappears for practical purposes (in that the atomic level introspection lacks the goal-directed motivation we actually desire, and certain architectures are claimed to be inpenetrable in this regard, and some architectures certainly require analysis to provide high-level explanations, and some of these may not be able to carry out this analysis themselves).

The apparent difference between a person and a computer is thus that the former can introspect best at the highest and most abstract levels, and the latter can introspect easiest at the lowest and most atomic levels.

Or is your definition circular, so that introspection is defined to be something that only people can do? And as for your definition of infallibility, who is the judge? What is truth? (as Pilate asked) Psychologists and Linguists know very well that they cannot trust subject's intuitions - their introspection is anything but infallible, it's not even consistent.

David Powers

Date: Fri, 20 Aug 93 13:27:21 -0800 From: pat hayes

Selmer Bringsjord writes:

>sb> Now tell me what premises are false in these arguments,
>sb> or tell me to flesh out the arguments on the spot, in
>sb> plain English

OK. Here's the first one: To assume that there's nothing it's like to be a computing machine simply begs the question. How on earth would you know? I take it that you and I are, in fact, computing machines (as well as other kinds of machine). You assume that we know what it's like on the inside: ergo, there IS something it's like to be a computing machine on the inside. I should point out that this phrase 'like on the inside' is quite meaningless to me, but that's OK, the logic of the argument doesn't depend on its having meaning. Now of course this begs the question, but no more than you do.

This is an example of a class of anti-AI arguments of the general form 'People have mysterious property X. Automata don't have X. Therefore people aren't automata." The difficult premise is always the second one, since it is never possible to establish this, and the argument might just as well be taken in modus tollens form to establish that, since people (obviously) are automata, automata do have property X. This conclusion is usually just taken to be OBVIOUSLY false (and sometimes offensive by some).

The second case is worse, since the premise is just obviously false. People's capacity to introspect is quite startlingly bad. We have known this since Freud; but without getting psychoanalytical, we are all familiar with such ordinary stuff as lapses and reconstructions of memory, having information on the tip of the tongue, not being able to locate a pain, being surprised at having little skills we weren't familiar with, finding ourselves getting irritated for no conscious reason, etc. etc.. We are hopeless introspectors: I think it likely that machines could be made to be much better at this than we are.

Selmer also gives us an outline of his argument from free will. This uses the concept of 'iterative agent causation', and concludes that this is true from a premise concerned with moral responsibility. Even if we assume the truth of it, however, his conclusion doesn't follow. The proposition is that people, not events, bring about certain events, and bring about their bringing about of those events. Let's accept that, whatever it means. He concludes that people aren't automata. Why?

Surely automata might also bring about certain events, and bring about their bringing about of those events. That seems quite consistent, and clearly allows people to be automata. One might try to argue that the kinds of events that people bring about are different from those that automata do, or some such, but this is clearly a hard line to hoe: and anyway, Selmer at this point gives the following reason: "if people can enter an infinite amount of states in a finite time, were they automata, they would be able to solve unsolvable computational problems." This seems to have absolutely nothing to do with the rest of the argument, which does not even mention numbers of states. I might add that the premise "were they automata" here is unnecessary, since if people can enter an infinite number of states in finite time then they could perform physically impossible feats no matter what they were, which suggests - if it needed suggesting - that they probably cannot do this.

So, my conclusion is that Selmer's arguments are, in spite of the hopeful hyperbole about their logical solidity, quite unable to establish his conclusions.

Pat Hayes

From: Selmer Bringsjord Date: Fri, 20 Aug 93 18:27:48 -0400

Pat Hayes writes:

>ph> To assume that there's nothing it's like to
>ph> be a computing machine simply begs the question.

The proposition that "there's nothing it's like to be a computing machine" isn't a premise *in the argument of Chapter I.* This proposition is there unpacked, and established by way of a separate argument.

>ph> How on earth would you know? I take it that you and I are, in fact,
>ph> computing machines (as well as other kinds of machine).

That's an interesting move. I shall have to remember to advise my religious friends to make a parallel move when debating atheists: Begin by assuming that God exists. All of "ambitious" AI/Cog Sci hinges on the claim that human persons are computing machines.

>ph> This is an example of a class of anti-AI arguments of the general
>ph> form 'People have mysterious property X. Automata don't have X.
>ph> Therefore people aren't automata."

Where, pray tell, did I say the property in question (being such that one can experience the raw feel of meeting a long lost friend, e.g.) is mysterious? In WRC&CB I point out that such properties undergird a normal appreciation of creative text. Your position would then be that appreciating a short story is mysterious. I won't even complete the *reductio*.

>ph> The difficult premise is always
>ph> the second one, since it is never possible to establish this, and the
>ph> argument might just as well be taken in modus tollens form to
>ph> establish that, since people (obviously) are automata, automata do

>ph> have property X. This conclusion is usually just taken to be
>ph> OBVIOUSLY false (and sometimes offensive by some).

Can't parse this.

>ph> The second case is worse, since the premise is just obviously false.
>ph> People's capacity to introspect is quite startlingly bad. We have
>ph> known this since Freud; but without getting psychoanalytical, we
>ph> are all familiar with such ordinary stuff as lapses and
>ph> reconstructions of memory, having information on the tip of the
>ph> tongue, not being able to locate a pain, being surprised at having
>ph> little skills we weren't familiar with, finding ourselves getting
>ph> irritated for no conscious reason, etc. etc.. We are hopeless
>ph> introspectors: I think it likely that machines could be made to be
>ph> much better at this than we are.

I know the literature on introspection well. I know all the attacks on introspection from case histories a la Sacks, insanity, weakness of the will, drug use, RT studies galore in cog psych, David Hume, Paul Churchland spy-behind-enemy-line cases, and on and on *ad nauseum*. In *What Robots Can & Can't Be* I defend the proposition that human persons enjoy what I call 'hyper-weak incorrigibilism.' Here's a relevant snippet from the start of the relevant chapter (IX) from WRC&CB:

The aspect of incorrigibilism around which the argument in this chapter revolves is what I have coined hyper-weak incorrigibilism, the view, in short, that humans have, *with respect to a restricted class of properties*, the ability to ascertain infallibly, via introspection, whether they have these properties. This power, I claim, is beyond the reach of the sort of robot envisioned as the final product of [AI]. But it is a power, in the human case, which even philosophers inclined to frown upon introspection have been unable to deny -- as we shall see.

It should be noted up front that I reject out of hand simple- minded, garden-variety versions of incorrigibilism. A sanguine view about introspection, according to which it is infallible across the board, is, to put it bluntly, something no one in their right mind buys anymore. And as far as I can tell, IUm still in my right mind. {p. 329)

>ph> Selmer also gives us an outline of his argument from free will.
>ph> This uses the concept of 'iterative agent causation', and concludes
>ph> that this is true from a premise concerned with moral
>ph> responsibility. Even if we assume the truth of it, however, his
>ph> conclusion doesn't follow. The proposition is that people, not
>ph> events, bring about certain events, and bring about their bringing
>ph> about of those events. Lets accept that, whatever it means. He
>ph> concludes that people aren't automata. Why? Surely automata
>ph> might also bring about certain events, and bring about their
>ph> bringing about of those events.

And bring about the bringing about of the bringing about of those events, *ad infinitum*? Did you catch the three dots?

>ph> That seems quite consistent, and
>ph> clearly allows people to be automata. One might try to argue that
>ph> the kinds of events that people bring about are different from
>ph> those that automata do, or some such, but this is clearly a hard
>ph> line to hoe: and anyway, Selmer at this point gives the following
>ph> reason: "if people can enter an infinite amount of states in a finite
>ph> time, were they automata, they would be able to solve unsolvable
>ph> computational problems". This seems to have absolutely nothing
>ph> to do with the rest of the argument, which does not even mention
>ph> numbers of states.

Of course, what's in the background is the argument of which this was an *outline*. The states in question correspond to bringing about certain mental events. The idea is that we are forced to countenance something which looks massively counterintuitive (that persons enter an infinite number of mental states in a finite amount of time). Turing machines allowed to do that could, e.g., solve the busy beaver function.

>ph> I might add that the premise "were they
>ph> automata" here is unnecessary, since if people can enter an
>ph> infinite number of states in finite time then they could perform
>ph> physically impossible feats no matter what they were, which
>ph> suggests - if it needed suggesting - that they probably cannot do
>ph> this.

Iterative agent causation has only been maintained by thinkers who see no other way to make room for freedom (e.g., Roderick Chisholm, Richard Taylor). Since I know, if I know anything, that whether or not I have pizza tonight is up to me, and I also know that determinism, indeterminism and normal agent causation imply that whether or not I have pizza tonight *isn't* up to me, something like iterative agent causation comes on stage.

How does it follow from "people can enter an infinite number of mental states in a finite amount of time" that "people can perform physically impossible feats"? Just curious.

>ph> So, my conclusion is that Selmer's arguments are, in spite of his
>ph> hopeful hyperbole their logical solidity, quite unable to
>ph> establish his conclusions.

Ya have to read the book, Pat! I'm trying to convince Stevan -- in the commentary in question -- that his *methodology* begs the question against an approach dating back to at least Plato. I'm not trying to zap off a few argument sketches in the wild hope that they'll be compelling. On the other hand, I keep hoping you'll come round. If you get out now, you have plenty of time to invest in positions on the mind which won't be, in a few years, carcasses littered along the path of intellectual history. Repent! Let your descendants look back and smile at your wise change of direction, made as the 20th century arrived.

Selmer Bringsjord

Date: Fri, 13 Aug 93 13:04 BST From: ronc@cogs.susx.ac.uk (Ron Chrisley)

Stevan Harnad wrote:

>sh> ON IMPLEMENTATION-DEPENDENCE AND COMPUTATION-INDEPENDENCE

>sh> Ron Chrisley suggests that the thesis that "Cognition is
>sh> Computation" is nonvacuous even if every physical process is
>sh> computation, as long as COGNITIVE computation can be shown to be a
>sh> special KIND of computation. (He does not, of course, suggest what
>sh> that special kind of computation might actually be, for here he is
>sh> only trying to establish that such a thesis is tenable.)

No, even if cognitive computation *doesn't* turn out to be a special kind of computation, that would just mean the "Cognition is Computation" ("C is C") claim would be *false*, not vacuous.

>sh> Ron does mention (and to a certain extent equivocates on) one
>sh> difference that may indeed distinguish two different kinds of
>sh> computation: the "implementation-DEpendent" kind (again, not
>sh> defined or described, just alluded to) and the usual,
>sh> implementation-INdependent kind. Ron thinks cognition may turn
>sh> out to be implementation-DEpendent cognition.

Just to make sure: the "implementation-INdependent vs implementation-DEpendent" distinction is not meant to be closely related to the "cognitive computation vs non-cognitive computation" distinction. It's just that I think the non-vacuity of ubiquitous computation holds whether or not the best account of computation is implementation-dependent. From now on, I'll make it clear which notion of computation I am talking about by using II for implementation-independent computation, and ID for implementation-dependent computation.

By ID computation, I mean an account of computation that sometimes assigns different computational properties to systems that differ *only* in what is classically (II) thought of as their implementations, and *not* in their traditional computational properties. Thus, ID computational properties (at least sometimes) cut finer than do II properties.

>sh> In writing about physical systems that science has NOT so far found
>sh> it useful to describe, study or explain computationally (the sun,
>sh> for example), Ron notes that they are nevertheless computational
>sh> systems,

No. I'm just being generous to my opponent when I consent to the idea of a computational description of the sun. *IF* computation is ubiquitous, then, ex hypothesi, the sun (and everything else) has a computational description. I want to show that even in such a case, the claim that cognition is computation is non-trivial.

(Now it might be the case that there are physical systems for which we can find *no* computational description. I doubt it, but even if it were the case, it would be a mistake to think that this is the way to save the "cognition is computation" claim from its supposed vacuousness.)

What was the point, then, in mentioning the sun? Well, suppose the sun has a computational description. Even so, the claim "stellar dynamics is computation" seems false. That is, it seems very likely that there is no computational kind that picks out all and only stars. Any computational description that a star meets could, no doubt, be instantiated by a system which we would not want to call a star. The best explanations of what's going on in stars are not computational. The computational description does not help us explain what we want to explain about stars.

Just as the claim that "stellar dynamics is computation" can be false even if everything has a computational description, so too the claim that cognition is computation might be false, even if everything has a computational description.

>sh> and in some sense "multiply realizable" (I'm not sure whether he
>sh> means this to be synonymous with implementation-independent -- I
>sh> don't think there's any cat that can only be skinned ONE way, but I
>sh> don't think that's quite what's ordinarily meant by
>sh> "implementation-independence" in the computational context,
>sh> otherwise that term too risks becoming so general as to become
>sh> vacuous.)

I assume that even ID computation can be multiply realized; even though an ID computational property may only hold of *some* II-computationally-equivalent systems, there are going to be many possible systems that *do* meet any implementation criteria.

(I was going to say that even on an ID account, any two qualitatively identical physical systems will necessarily have the same computational properties, but that would obscure the possibility of ED -- environment-dependent -- computation, which can assign different computational properties to identical systems, as long as their environments are suitably different.)

>sh> I'll lay my own cards on the table, though: The only
>sh> implementation-independence *I* think is relevant here is the kind
>sh> that a computer program has from the computer that is implementing
>sh> it. Is that the kind Ron thinks the sun has? (I mean the sun in our
>sh> solar system here, not the workstation by that name, of course!)

Well, it depends on which notion of computation we are dealing with at this point. Let's stick to formal, II computation for the time being. Then, yes. *IF* everything has a computational description, then the sun does, and the description that it has could be realized by many things (since we are right now talking about II computation).

>sh> If so, then a computer simulation of the sun -- a physical symbol
>sh> system implementing the sun's COMPUTATIONAL description, in other
>sh> words, a computer running the sun's computer program and hence
>sh> systematically interpretable as the sun, a virtual sun -- would
>sh> have to be hot (VERY hot, and not just symbols systematically
>sh> interpretable as if they were hot). We don't even need Searle to see
>sh> that THAT's not likely to happen (because you can FEEL the absence
>sh> of heat, but you can't be so sure about the absence of mind). So
>sh> whatever kind of "implementation-independence" the sun may have,
>sh> it's not the kind we need here, for the purposes of the "cognition

>sh> is computation" thesis.

No, a computer running the sun’s program would *not* have to be very hot. I would only believe that if I thought that *being a sun* was a computational property. But it is not: it is possible that X has a computational description, and that X has property P, and yet P is not a computational property. This would be shown to be the case if there were some systems that also had X’s computational description, yet lacked P.

I agree (with Searle) that re-instantiating the computational properties of X does not guarantee re-instantiating all the properties of X (this is obvious, and is required to even make sense of the notion of implementing the same program in different hardware). So a computer simulation of a sun is not (typically) a sun. So *being a sun* would not be a computational property, *even if everything had a computational description*. So it is possible that "being a cognizer is a computational property" is also false, even if everything has a computational description. So computational ubiquity does not render the "cognition is computation" claim vacuous.

>sh> So suppose we give up on that kind of software/hardware
>sh> implementation-independence and settle for
>sh> "implementation-DEpendent" computation -- whatever that is, for it
>sh> sounds as if spelling out the nature of that dependence will turn
>sh> out to be as essential to a description of such a "computational"
>sh> system as the computational description itself.

Yes, it is intended to! ID computation cuts finer than II computation, and says that what II computational theory *thought* to be computationally irrelevant (time, physical constitution, connection with subject matter, etc.) is actually very relevant, and constitutive of certain computational properties.

>sh> Indeed, it sounds as if the dependence-story, unlike the
>sh> computation-story, will turn out to be mostly physics in that case.

I find that very unlikely, given the intentional nature of cognition. The expression, in the non-intentional language of physics, of a temporal constraint in an ID computational theory (especially of a cognitive system) would be as wildly disjunctive as a physical expression of an II computational constraint.

>sh> I mean, I suppose flying is an implementation-dependent sort of
>sh> computation too, and that a plane is, in a sense, just a highly
>sh> implementation-dependent computer. The only trouble is that all the
>sh> RELEVANT facts about planes and flying are in the physics
>sh> (aeronautical engineering, actually) of that dependency, rather than
>sh> the computation!

That last bit is right. That’s why flying is *not* best thought of as computation, be it II or ID. And that’s why the claim that *cognition* is computation is neither incoherent nor vacuous: because it can be false.

>sh> So if we open up the Pandora's box of implementation-dependence, is
>sh> there not the risk that the "Cognition is (implementation-dependent)
>sh> Computation" thesis would suffer the same fate as a "Flying is
>sh> (implementation-dependent) Computation" thesis?

Yes. It could be false. That risk buys non-vacuity.

>sh> Now to the play-by-play:

rc> Date: Tue, 15 Jun 93 18:35 BST rc> From: ronc@cogs.susx.ac.uk (Ron Chrisley)

rc> I said C has to be a *natural* and *non-question-begging* class rc> of computation. Simply *defining* what the brain does as rc> "cognitive computation" is not going to get one anywhere. One rc> has to show that there is a class, naturally expressible in rc> terms of computational concepts, that includes brains and all rc> other cognizing physical systems, but leaves out stars, stones, rc> etc. Only then will one be justified in claiming "cognition is rc> computation" in any non-vacuous sense. If the best one can do, rc> when using computational concepts, is to find some wildly rc> disjunctive statement of all the systems that are cognizers, rc> then that would suggest that cognition is *not* computation. So rc> the claim is not vacuous, but contentious.

>sh> It's not coincidental that the "Is Cognition Computation?" and the
>sh> "What is Computation?" discussion threads are linked, because it's
>sh> critical to get it straight what it is that we are affirming or denying
>sh> when we say Cognition Is/Isn't Computation. I was content to assume that
>sh> what was at issue was implementation-independent symbol manipulation,
>sh> but now, with the introduction of TMs (Turing Machines) in place of
>sh> UTMs (Universal Turing Machines) it's being suggested that that's not
>sh> the issue after all.

None of what I've said even mentions the TM/UTM distinction. I don't think it relies on it.

>sh> It seems to me that although your burden is not to actually produce
>sh> the RIGHT theory of what is special about that subset of computation
>sh> that is cognitive, you do have to give us some idea of the KIND of
>sh> thing it might be.

I'm not sure that I have this burden in order to discredit the claim that computational ubiquity renders the C is C claim vacuous. But I'll give it a shot.

>sh> So let's look more closely at your implementation-dependent TM
>sh> theory of mind.

Again, "ID vs II" is not supposed to equal "cognitive vs non-cognitive computation".

>sh> Here's an important question: Would the UTM simulation of the
>sh> right TM (the one that had mental states) have mental states?

It depends on whether your computational account of cognition is II or ID. I'm trying to show that II or ID, the C is C claim is not made vacuous by computational ubiquity.

>sh> If so, we're back to Searle.

Yes, if your account of computation is II, then you need a response to Searle (many are available) if you want the C is C claim to be *true*. Else, it is *false*. That would support, not defeat, my claims here. I'm not (primarily) arguing for the C is C claim; I'm arguing that it is non-vacuous, even if computation is ubiquitous.

>sh> If not, I'd like to know why not, since computational equivalence is
>sh> supposed to be the pertinent INVARIANT that holds all these
>sh> computational descriptions together. I mean, without the
>sh> computational equivalence, isn't it back to physics again?

But if you subscribe to an ID theory of computation, you explicitly *deny* that TM equivalence guarantees computational equivalence! Why should we think that just because an alternative account of computation cuts more finely than traditional computational categories do, that we have left the computational/intentional and landed in the mere physical?

>sh> In fact, computationalism was supposed to show us the
>sh> DIFFERENCE between physical systems like the sun and physical
>sh> systems like the brain (or other cognitive systems, if
>sh> any), and that difference was supposed to be that the brain's
>sh> (and other cognitive systems') cognitive function, UNLIKE the
>sh> sun's solar function, was implementation-independent - i.e.,
>sh> differential-equation-independent - because cognition really
>sh> WAS just (a kind of) computation; hence every implementation
>sh> of that computation would be cognitive, i.e. MENTAL.

Yes, this is still the basic idea. What is being proposed here (and this is getting more like a defense of ID computation, rather than defending the bite of the C is C claim) is that what II computational theory would take to be two implementations of the same computation are instead implementations of two different computations.

rc> That might have been how many people thought computation was rc> relevant to cognitive science, but then one can take what I say rc> here to be a different proposal. I think both the sun and the rc> brain can be looked at as performing some computation. So what's rc> special about cognition is not that it is realized in a system rc> that can be looked at computationally.

>sh> (I don't mean to pick on your phraseology, but that last sentence
>sh> sounds like a denial of the Cog=Comp thesis right there...)

It shouldn't look that way. Database programs, formal system games, etc. are realized in systems that can be looked at computationally, yet they are not cognitive. The C is C claim does *require* that cognition be realized in systems that can be looked at computationally. But it does not imply that this is what makes cognition special. In fact, it might be that *everything* is realized in systems that can be looked at computationally, yet the C is C claim might be false. Or at least that's what I've argued.

>sh> But of course you are here adverting to the fact that it's going to
>sh> turn out to be a special KIND of computation. Can you be more
>sh> specific, perhaps give examples of other systems that have been
>sh> taken to be natural kinds because they turned out to be special
>sh> kinds of computational systems?

I think what you meant to say was, "perhaps give examples of other phenomena that have been taken to be essentially computational because there turned out to be natural kinds of computational systems." Because that's what I take the C is C claim to be saying: cognition is computation because it can be naturally characterized in computational terms, and perhaps (scientifically) only in those terms. I don't have many other examples, but perhaps something like *sorting* is essentially computational in this sense.

>sh> And can you suggest what lines the
>sh> specialness might take in the case of cognition?

Of course I can't give *the* answer, because cognitive scientists haven't found one (yet?). Well, current orthodoxy says that cognitive systems are those systems that have memory, can learn in these particular ways, can reason common-sensically in these ways, have some type of perception/action link with the world, etc. (Note that connectionists usually don't try to come up with theories of *Cognition* in general, but rather models of how mammals might do it. Most don't claim that *only* something which realizes a connectionist network can be cognizing; but some do). The hypothesis is that the characterizations that will allow us to distinguish systems that have these essential cognitive features from those that do not will be *computational* characterizations. This might be false; it certainly isn't vacuously true, even if everything has a computational characterization. It might be the case that although everything has a computational characterization, any proposed set of distinctions will fail because either: a) the distinctions are not expessible computationally (e.g. "something must have a soul to be cognitive"); b) the distinctions do not actually divide up the world into the cognitive and non-cognitive (e.g. "something must instantiate Interlisp-D in order to be cognitive"; [probably many cognitive things *don't* instantiate any dialect of LISP!]); or c) both a and b are true. (The examples are deliberately unreasonable and extreme, in order to distinguish the possibilities clearly).

rc> Nor is the multiple realizability feature significant here. Once rc> one admits that there is a computational description of what the rc> sun is doing, then one ipso facto admits that in some sense, rc> what the sun is doing is multiply realizable too. So that's not rc> what is so computationally special about cognition.

>sh> As I said, multiple-realizability is not quite the same as
>sh> implementation-independence. There are, for example, many different
>sh> ways to transduce light, some natural (the vertebrate retinal cone
>sh> or the invertebrate omatidium), some artificial (as in a bank-door's
>sh> photosensitive cell), but NONE of them are computational, nor do
>sh> any constitute a natural family of computationally equivalent systems
>sh> -- or if they do, the computational story is trivial. It's the physics
>sh> of light-transduction that's relevant.

That all seems right.

>sh> By way of contrast, all the things you can reconfigure a computer to
>sh> DO by changing its software DO share interesting properties, and the
>sh> properties are computational ones: The same software can be run on
>sh> radically different forms of hardware yet it would still be
>sh> performing the same computation. THAT was the kind of
>sh> multiple-realizability that I THOUGHT was relevant to what
>sh> computation is and what Cog=Comp claims.

Yes. ID computation just puts greater restriction on what can count as an implementation of a given computation. So there will still be multiple realizability, of a more restricted sort.

>sh> By way of contrast, note that the number of ways you can reconfigure
>sh> a UTM like a digital computer to implement different programs does
>sh> NOT include a way to reconfigure it into an optical transducer, a
>sh> plane, or a sun.

On all II theories of computation, and all reasonable theories of ID computation, that's right. That shows that those entities are not, essentially, computational. That just shows that the claims of the form "x is computation" are not vacuous, even if everything has a computational description.

>sh> For that, the "reconfiguring" would have to be more radical than
>sh> merely computational: It would have to be physical. (And that's why
>sh> I think implementation-DEpendent "computation" is a non-starter.)

I've already questioned, above, this assumption of yours, that if a difference isn't a formal computational difference, then it must be merely a physical difference.

rc> What's special is this: there is no *natural*, rc> *non-wildly-disjunctive* way to distinguish white dwarf stars rc> from red giant stars by appealing to the computational systems rc> they instantiate. The "cognition is computation claim", however, rc> is claiming that there *is* a natural way to distinguish rc> cognizing systems from non-cognizing ones. And that natural way rc> is not using the concepts of neurophysiology, or molecular rc> chemistry, but the concepts of computational theory. This is a rc> substantive claim; it could very well be false. It certainly rc> isn't vacuously true (note that it is not true for the case of rc> stars); and its bite is not threatened even if everything has a rc> computational characterization.

>sh> But, as I said, I UNDERSTOOD the claim when I took computation to be
>sh> implementation-independent symbol-manipulation.

OK; you seem to agree here that if computation is II, then one has the non-vacuity of the cognition is computation claim.

>sh> But with implementation-DEpendent TMs I no longer even know what's
>sh> at issue...

I never said ID TM's. On my terminology, that is an oxymoron, since TM's are essentially II.

>sh> "Not wildly disjunctive" just isn't a positive enough
>sh> characterization to give me an inkling. Do you have examples, or
>sh> (relevant) analogies?

Imagine two systems that are formally computationally identical, yet their environments differ. Often, if these systems are sufficiently interactive with their environments (e.g. feedback-oriented), we will find it useful to classify them into different informational/computational states. What would have to be identical computational systems from an II point of view can be seen, from an ID (or ED, see above) point of view, distinct.

(I was trying to write a paper on this, but pretty much gave up when I found a paper that pretty well expresses my intuitions here, and argues (successfully, I think) against people like Fodor, who think that non-externally individuated states are the only scientific way to go. If externally individuated states are scientifically respectable, I suspect they will be very useful in a real-world computer science. The paper in question is by Raimo Tuomela, "Solipsism and explanation in psychology", Phil of Sci 56, 1989, esp p 39, passim.)

>sh> Let me put it even more simply: It's clear that some subset of
>sh> computer programs is the subset that can do, say, addition. Let's
>sh> suppose that this subset is "not wildly disjunctive." It is then an
>sh> equivalence class, of which we can say, with confidence, that every
>sh> implementation of those computer programs will be doing addition.
>sh> Now all you need is a similar story to be told about thinking: Find
>sh> the right ("nondisjunctive") subset of computer programs, and then
>sh> every implementation of them will be thinking.

Yes, if your theory of computation is II.

>sh> But now you seem to be saying that NOT every implementation of them
>sh> will be thinking, because the programs are implementation DEpendent.

No, I only consider this possibility, to show that ubiquity of computation is not a problem for the C is C claim even on an ID view of computation. On such a view, there can be systems that are II computational (TM) equivalent, that are not cognitively equivalent. But ID theorists could still hold that *ID* equivalent systems must be cognitively equivalent.

>sh> So what does that leave of the claim that there is a nontrivial
>sh> COMPUTATIONAL equivalence there to speak of at all?

I hope I've answered this.

>sh> Remember, if we go back to the sun, the scientific story there is
>sh> thermodynamics, electromagnetism, etc. It's not in any interesting
>sh> sense a computational story. Solar physics is not a branch of
>sh> computer science. No one is espousing a "Solar Dynamics =
>sh> Computation" hypothesis.

Exactly. So that hypothesis is not vacuous, it is false. Same with the C is C hypothesis, only we're not so sure it's false. But it is, at least, non-vacuous.

>sh> All of physics is (approximately) COMPUTABLE, but that does not mean
>sh> that physical processes are COMPUTATIONAL.

Exactly. So even if every system has a computational description, we have the non-vacuity of claims of the form "X is Computation", which is all I was arguing for.

>sh> And as far as I can tell, the most direct CARRIER of that
>sh> dissociation is the fact that physics is not
>sh> implementation-independent. So computational equivalence has a
>sh> hollow ring to it when you are trying to explain the physics. The
>sh> burden is to show why exactly the same thing is not true when it
>sh> comes to explaining thinking.

Yes. Again, there are many things which might have the same computational description as the sun, but that does not mean it will have all of the sun's essential properties.

rc> Now if everything can have *every* computational rc> characterization, then the claim might be in danger of being rc> content-free. But that's why I took the time to rebut Putnam's rc> (and in some sense, Searle's) arguments for that rc> universal-realization view.

>sh> As it happens, I never accepted the "everything is EVERY
>sh> computation" view (for the cryptographic reasons I adduced earlier
>sh> in this discussion). But I think "everything is SOME
>sh> [implementation-independent OR implementation-dependent]
>sh> computation" is just as empty and unhelpful as a basis for a
>sh> cognition-specific thesis, for that's just the Church-Turing Thesis,
>sh> which is a formal thesis about what "computation" is and has NOTHING
>sh> to do with mental states or what they are or aren't.

I'm not sure if the ubiquity of computation is the same as the Church-Turing thesis. But if you deny the ubiquity of computation (deny that everything has a computational description), then you can *ignore* my claim! Because my claim was conditional: Even *IF* everything has a computational description, the C is C claim is not vacuous.

rc> I agree that Searle was trying to show that the "cognition is rc> computation" claim is false. But his argument applies (although rc> I don't feel it succeeds) to my construal of the "C is C" claim. rc> He was trying to show that there was no such class of rc> computation that characterizes cognizers, since he could rc> instantiate one of the computations in any proposed class and rc> not have the requisite cognitive properties.

>sh> Yes, but Searle's argument works only for computation construed as
>sh> implementation-INdependent symbol manipulation. If some other sense
>sh> of computation is at issue here, his argument may well fail, but
>sh> then I don't know what would be at issue in its place.
>sh>
>sh> Indeed, Searle's argument fails immediately if anyone wants to say
>sh> (as, say, Pat Hayes does): Cognition has to be implemented the

>sh> "right way" and Searle's implementation is not the right one. But to
>sh> save THAT from amounting to just special pleading in the same sense
>sh> that, say, a "wild disjunction" would be, one has to face the
>sh> problem of how to distinguish the "right" from the "wrong"
>sh> implementation without shifting the scientific substance of the
>sh> explanation of cognition to the implementational details rather than
>sh> the computation!

If you don't like ID computation, ignore it. I have argued that the C is C claim is non-vacuous on the II notion as well.

(To address the Chinese Room argument very quickly: I agree that this spelling out of ID computation should be done. But as long as it is even a mere possibility, then Searle's argument fails. Searle is the one making the strong claims; he has not *shown* that these claims are true until he shows us that there is no possibility of ID computation.)

>sh> Again, to put it ever so briefly: Implementation-DEpendent
>sh> "computation" would indeed be immune to Searle, but look at the
>sh> price: Cognition is now not just the right computation, but the
>sh> right implementation of that computation -- and then the rest is
>sh> just an arbitrary squabble about proportions (computations/physics).

I have, several times, said why this is not an accurate view of the ID situation. On the ID view, one *does* only have to look at what is computationally relevant in order to determine cognitive properties; it is just that the ID theorist considers *more* distinctions to be computationally relevant than does the II theorist. It could be that these extra distinctions *are* non-computational, non-intentional, merely physical. That would imply that the ID view of computation is incorrect. But we don't know that it is incorrect right now. And even if it is incorrect, the ubiquity of computation does not make the C is C claim vacuous!

>sh> Now the implementation-independence of the symbolic level of
>sh> description is clearly essential to the success of Searle's
>sh> argument, but it seems to me that it's equally essential for a
>sh> SUBSTANTIVE version of the "Cognition Is Computation"
>sh> hypothesis (so-called "computationalism"). It is not
>sh> panpsychism that would make the hypothesis vacuous if
>sh> everything were computation; it would be the failure of
>sh> "computationalism" to have picked out a natural kind (your "C").

rc> As computational theory stands now, it is pretty rc> implementation-independent. Such a purely formal theory may or rc> may not (I go with the latter) be the best account of rc> computation.

>sh> It's at this point that I sense some equivocation. We are now to
>sh> envision not only a still unspecified "special" KIND of computation
>sh> that is peculiar to (and sufficient for) mental states, but we must
>sh> also imagine a new SENSE of computation, no longer the old
>sh> implementation-independent kind on which the whole formal theory was
>sh> built. My grip on this inchoate "computation" is loosening by the
>sh> minute...

No equivocation, as I'm sure you now see. As for losing grip: fine. Go back to good-old computation, and you still have the non-vacuity of the C is C claim.

rc> But even if our notion of computation were to change to a more rc> "implementation-dependent" notion (although I suspect we rc> wouldn't think of it as more "implementation-dependent" once we rc> accepted it), I don't see why the "C is C" claim would be in rc> danger of vacuity. It would be even stronger than before, rc> right? But perhaps you just meant that it would be false, since rc> it would rule out cognizers made of different stuff? That's just rc> a guess that the "C is C" claim is false: that the natural kinds rc> of computation do not line up with cognition. That's not an rc> argument.

>sh> I couldn't be suggesting that such a claim was false, since, as I
>sh> said, I've lost my grip on what the claim is claiming!

Again, don't worry. I shouldn't have mentioned ID computation at all, it seems! I guess the only reason to mention it is if: 1) you think computation is ubiquitous; 2) you think a Hayes-style response is the only way to respond to the Chinese Room; 3) you want the C is C claim to be non-vacuous. Since you, Stevan, don't meet at least 2 of these conditions, it is a bit irrelevant.

>sh> "Stuff" had nothing to do with the old Cog=Comp thesis, since it was
>sh> implementation-independent. And I could quite consistently manage to
>sh> be an anticomputationalist toward this old form of computationalism
>sh> (because of the symbol grounding problem)

But could you maintain that the C is C claim is *vacuous*, or merely *false*?

>sh> without for a minute denying that minds could be realized in
>sh> multiple ways (just as optical transducers can); in fact, that's
>sh> what my Total Turing Test (T3) banks on. But SYNTHETIC alternative
>sh> realizations (made out of different, but T3-equivalent stuff) are
>sh> not the same as VIRTUAL alternative realizations, which is what a
>sh> pure symbol system would be. Besides, a symbol-cruncher alone could
>sh> not pass T3 because it lacks sensorimotor transducers --
>sh> significantly, the only part of a robot that you can't have a
>sh> virtual stand-in for.
>sh>
>sh> But never mind that: Why would a Cog=Comp thesis involving
>sh> "implementation-dependent computation" be stronger than one
>sh> involving implementation-independent computation? It sounds more
>sh> like a weaker one, for, as I keep hinting, there is the risk that in
>sh> that case the relevant functions are in the DEPENDENCY, i.e., in the
>sh> physics (e.g., the transduction) rather than the computation.

I meant stronger in the logical sense. An ID notion of computation will mean that even fewer systems are cognizers, which is even further away from the panpsychism that C is C, plus computational ubiquity, supposedly entails.

>sh> For we could easily have had a "computationalist" thesis for
>sh> solar heat too, something that said that heat is really just a
>sh> computational property; the only thing that (fortunately)
>sh> BLOCKS that, is the fact that heat (reminder: that stuff that
>sh> makes real ice melt) is NOT implementation-independent, and
>sh> hence a computational sun, a "virtual sun" is not really hot.

rc> Right, but the computation that any particular hot thing rc> realizes *is* implementation-independent. The question is rc> whether that computation is central to the phenomenon of heat, rc> or whether it is accidental, irrelevant to the thing *qua* hot rc> thing.

>sh> Well, apart from the fact that thermodynamics, the science of heat,
>sh> shall we say, is not accustomed to thinking of itself as a
>sh> computational science, surely the ESSENTIAL thing about heat is
>sh> whatever "being hot" is, and that's exactly what virtual heat lacks.

Agreed. And the falsity of such a thesis ensures its non-vacuity.

>sh> If this is not obvious, think of a virtual plane in a virtual world:
>sh> It may be computationally equivalent to a real plane, but it can't
>sh> fly! I wouldn't describe the computational equivalence between the
>sh> real and virtual plane as accidental, just as insufficient -- if
>sh> what we wanted was something that FLEW!
>sh>
>sh> And EXACTLY the same is true in the case of wanting something that
>sh> really has mental states -- at least with the old candidate:
>sh> implementation-independent symbol manipulation (which could yield
>sh> only a VIRTUAL mind, not a real one). But I have no idea how to get
>sh> a grip on an implementation-DEPENDENT candidate. I mean, what am I
>sh> to suppose as the TM in question? There's (1) real me. I have real
>sh> mental states. There's (2) virtual "me" in a virtual world in the
>sh> pure symbol cruncher. It's computationally equivalent to me, but for
>sh> Searlean reasons and because of the symbol grounding problem, I
>sh> don't believe for a minute that it has a mind.

OK, so on your view the C is C claim is false. I'm not so sure, but at least we can agree that it is not vacuous.

>sh> But that's not what's at issue here. We should now think of a third
>sh> entity: (3) A TM performing implementation-DEpendent computations. I
>sh> can't imagine what to imagine!

Well, it's an instantiated TM, but it may have a different environment, etiology, and/or temporal character than you do. If so, it may be computationally distinct from you, even though it is TM-equivalent.

>sh> If it's a robot that's T3-indistinguishable, I'm already ready to
>sh> accept it as a cognizing cousin, with or without a computational
>sh> story (it could all be a transducer story -- or even a HEAT story,

>sh> for that matter, in which case real heat could be essential in BOTH
>sh> cases).

An ID computational theory could lump together as computationally equivalent systems that are T3-distinguishable.

>sh> But what am I to imagine wanting to DENY here, if I wanted to deny
>sh> this new form of computationalism, with TMs and
>sh> implementation-DEpendence instead of UTMs and
>sh> implementation-INdependence?

If we've changed topics from the issue of vacuity/ubiquity to "is the C is C claim true, under an ID notion of computation?", then all you have to imagine is a theory of computation that cuts finer than formal accounts do nowadays.

rc> If you want to avoid begging the question "is heat rc> computation?", then you have to allow that heat *might* be rc> implementation-independent. Then you notice that there is no rc> natural computational class into which all hot things fall, so rc> you reject the notion of "heat is computation" and thereby the rc> prospects for implementation-independence.

>sh> Ron, you've completely lost me. There's no sense of heat that I can
>sh> conjure up in which heat is computation, no matter how many ways it
>sh> can be realized.

Same here.

>sh> (Again: multiple-realizability is not the same as
>sh> implementation-independence.) I have no problem with synthetic heat,
>sh> but that still does not help me see heat as computation (the real
>sh> problem is VIRTUAL heat). And even if there is a nice, crisp
>sh> ("nondisjunctive") set of unique computational invariants that
>sh> characterize hot things and no others, I still don't see what it
>sh> would mean to say that heat was computation -- except if EVERY
>sh> implementation of the heat program were hot -- which is decidedly
>sh> not true (because virtual heat is not hot).

But every implementation of the heat program *would* be hot, if we grant your hypothesis that there is a computational characterization that captures hot things and no others.

>sh> (Also, in the above passage it sounds as if you are conceding that
>sh> computationality DOES call for implementation-INdependence after
>sh> all.)

That's because at that point in the discussion it was being assumed that computation is II.

>sh> In contrast, computationalism was supposed to have picked out
>sh> a natural kind, C, in the case of cognition: UNLIKE HEAT
>sh> ("thermal states"), mental states were supposed to be
>sh> implementation-independent symbolic states. (Pylyshyn, for
>sh> example, refers to this natural kind C as functions that

>sh> transpire above the level of the virtual architecture, rather
>sh> than at the level of the physical hardware of the
>sh> implementation; THAT's what's supposed to distinguish the
>sh> cognitive from the ordinary physical).

rc> No. *All* formal computations are implementation-independent.

>sh> Again, there seems to be some equivocation here. I thought you had
>sh> said you thought that cognition might be an implementation-DEpendent
>sh> form of computation earlier. Or is there now "formal computation"
>sh> and "computation simpliciter" to worry about?

No, there are two possible views of computation: II and ID. II is formal.

>sh> Entities seem to be multiplying and it sounds like it's all in the
>sh> service of saving an increasingly vague if not vacuous thesis...

rc> So it cannot be implementation-independence that distinguishes rc> the cognitive from the non-cognitive (now you see why I reminded rc> us all that "all cognition is computation" does not mean "all rc> computation is cognition"). There are many other phenomena whose rc> best scientific account is on a computational level of rc> abstraction (what goes on in IBM's, etc.; perhaps some economic rc> phenomena, et al). So the fact that a phenomenon is best rc> accounted for on the computational (implementation-independent) rc> level does not mean that it is cognitive. It means that it is rc> *computational*. The big claim is that cognition is a rc> computational phenomenon in this sense.

>sh> No, just as I have indicated that I of course realize that "all cog
>sh> is comp" does not imply "all comp is cog," I don't think that the
>sh> emphasis on the implementation-independence of cognition was enough
>sh> to uniquely characterize cognition, for there are of course plenty
>sh> of other implementation-independent computational phenomena. It had
>sh> more of a necessary- than a sufficient-condition flavor -- though of
>sh> course necessity was not at issue at all.

Then II is neither sufficient NOR necessary for cognition?

>sh> What Pylyshyn was saying was that mental states were unlike other
>sh> kinds of physical states, and were like other kinds of computational
>sh> states (including nonmental ones) in being IMPLEMENTATION-
>sh> INDEPENDENT. That wasn't SUFFICIENT to make every computation
>sh> mental, but it was sufficient to distance cognitive science from
>sh> certain other forms of physicalism, in particular, the kind that
>sh> looked for a hardware-level explanation of the mind.

But if you take Searle seriously, you must think Pylyshyn is wrong. So you should leave open the possibility of ID computation: a computation that cuts finer than does II computation (or use the Systems Reply -- oops!).

>sh> Now I had asked for examples. IBM is a bad one, because it's a
>sh> classical (approximation to a) UTM. Economic phenomena are bad
>sh> examples too, for of course we can model economic phenomena (just as
>sh> we can model solar systems), and all that means is that we can
>sh> predict and explain them computationally. I would have conceded at
>sh> once that you could predict and explain people and their thoughts
>sh> computationally too. I just don't think the computational oracle
>sh> actually THINKS in so doing, any more than the planetary oracle
>sh> moves (or the virtual plane flies or the virtual sun is hot).
>sh> "Economics" is such an abstract entity that I would not know what to
>sh> do with that; it's not a concrete entity like a person or a set of
>sh> planets. If you make it one, if you say you want to model "society"
>sh> computationally, then I'll say it's the same as the solar system
>sh> oracle. There's no people in the one, no planets in the other,
>sh> because such things are not implementation-independent.

Fair enough. Bad examples.

>sh> And it was on THIS VERY VIRTUE, the one that made
>sh> computationalism nonvacuous as a hypothesis, that it foundered
>sh> (because of Searle's argument, and the symbol grounding
>sh> problem).

rc> I accept that "implementation-dependent" (better: not purely rc> formal) notions of computation are probably the way to go (I've rc> argued as much on this list before), but I don't feel that rc> Searle compels me to do this.

>sh> Fine, but I have no idea what I am committing myself to or denying
>sh> if I accept or reject this new, "nonformal" form of
>sh> computationalism. I doubt if Searle would know either.

But in Chapter 9 of _The Rediscovery Of The Mind_ (p 209), Searle admits that there are people (e.g. Brian C. Smith of Xerox PARC) who are trying to come up with more embodied notions of computation that *do* place non-formal constraints on the "implementation" when taxonomizing, and that such notions of computation escape his vacuity argument.

>sh> Searle has, for example, never denied the possibility of synthetic
>sh> brains -- as long as they have the relevant "causal powers" of the
>sh> real brain. A pure symbol-cruncher, he has shown, does NOT have the
>sh> relevant causal powers. So if you tell him that all the systems that
>sh> DO have that causal power are computationally equivalent, he'll just
>sh> shrug, and say, fine, so they are, just as, perhaps, all stars are
>sh> computationally equivalent in some way. The relevant thing is that
>sh> they have the right causal powers AND IT'S NOT JUST COMPUTATION,
>sh> otherwise the virtual version -- which is, don't forget, likewise
>sh> computationally equivalent -- would have the causal powers too, and
>sh> it doesn't. Now that you've disavowed UTMs, pure formal syntax and
>sh> implementation-independence, this does not bother you, Ron; but just
>sh> what, exactly, does it leave you with?

It leaves me with the possibility of a computer science that takes seriously the environment, time, physical realization, and -- who knows -- perhaps history. Which is just as well, since that is the kind of computer science that some CS researchers are putting forward as necessary to explain today's complex information processing systems.

I think the primary dispute here has become: "how could one have an implementation-dependent notion of computation, given that computational properties are *defined* to be implementation-independent?" I reject that computational notions are defined to be II. To say so is to confuse the map with the territory. In discussing what computation is, my principal allegiance is to the phenomenon of computation, what goes on in IBM's, token rings, etc. If the best non-physical, intentional, representational account of that phenomenon individuates properties differently than does a formal account, then that just shows that computation isn't just formal symbol manipulation.

When I consider the C is C claim, I don't take it to be the claim that "Cognition is formal symbol manipulation, which may or may not be the best account of computational phenomena" but rather "Cognition is computation, and the best account of computation may or may not be in terms of formal symbol manipulation". That's why a notion of ID computation makes sense.

rc> The sociology of computer science certainly hasn't worked that rc> way. It just so happens that embedded and embodied notions are rc> essential to understand normal, ordinary computational systems rc> like token rings, CPU's hooked to robot arms, etc. But there is rc> a disagreement here: if computational theory goes embedded rc> (rejects implementation-independence), that doesn't mean it is rc> vacuous; just the opposite! It makes its range of application rc> even more restricted.

>sh> Symbol grounding is not just "embedded" symbol crunchers with
>sh> trivial add-on peripherals; but never mind.

That may be, but who mentioned symbol grounding? Not me.

>sh> I do agree that T3 exerts constraints on ALL models (whether
>sh> computational or, say, analog-transductive), constraints that T2,
>sh> imagination, and virtual worlds alone do not. I've already declared
>sh> that I'm ready to confer personhood on the winning T3 candidate, no
>sh> matter WHAT's going on inside it. The only thing that has been ruled
>sh> out, as far as I'm concerned, is a pure symbol cruncher.
>sh>
>sh> But since you never advocated a pure symbol cruncher, you would have
>sh> to say that Searle is right in advocating a full neuro-molecular
>sh> (T4) understanding of the brain, because the brain, like everything
>sh> else, is a computational system, and if "C is C" is right, Searle
>sh> will end up converging on the very same computational theory
>sh> everyone else does.

IF C is C, then the correct analysis of the brain will be at a level higher than the mere neuro-molecular. Just as an engineer can look at bits only, and not see the computational pattern that explains an IBM, so too Searle can insist on looking at neuro-molecular stuff and ignore the computational patterns in the brain.

>sh> My own guess is that going "embedded" or "implementation-dependent"
>sh> amounts to conceding that the cognition is in the physics rather
>sh> than just the computation. What shape that physics actually ends up
>sh> taking -- whether it is just a matter of hooking up the right
>sh> peripherals to a symbol cruncher in order to make the mental lights
>sh> go on or (as I suspect) there's rather more to it than that -- is
>sh> beside the point. The logical implication stands that without the
>sh> (shall we call it) "computation-independent" physics -- the RIGHT
>sh> (non-wildly disjunctive) physics -- there is no cognition, even if
>sh> the computation is "right."

Your *guess* might be right. But so might the *guess* that these implementational factors are themselves computational/intentional, and are not best understood with a mere physical vocabulary.

Surely there is little that a Martian and I need have *physically* in common for us to both be cognizers. It seems that the best level of analysis of our similarities would be above the physical, but below the formal (if the Chinese Room bothers you, for instance).

rc> My original claim (the negation of the one you made in your rc> initial response to Yee) was that *even if* everything has a rc> computational characterization, that does not make the rc> "computation is cognition" claim vacuous. I have given the rc> reasons above. Now if we have an implementation-dependent rc> computational theory, that does not mean that not everything rc> will have a computational characterization. It could just mean rc> that tokens that were of the same computational type in the rc> formal theory are now of distinct types in the embedded theory. rc> Nevertheless, despite such ubiquity of computation, there might rc> still be a natural computational kind which includes just those rc> things which are cognizers. Or there might not.

>sh> I don't know about types and tokens, but if there are two physical
>sh> systems that are both implementations of the very same formal
>sh> (computational) system and one of them is "right" and the other one
>sh> is "wrong," then it sounds as if the formal (computational) story is
>sh> either incorrect or incomplete.

In that case, ID computation suggests itself.

>sh> To resolve the ambiguity inherent in computationalism it is
>sh> therefore not enough to point out, as Ron has done, that the
>sh> "Cognition is Computation" thesis just claims that "Cognition is a
>sh> KIND of Computation"; for what it really claims is that "Cognition
>sh> is JUST a Kind of Computation."
>sh> And it's that "JUST" that I think implementation-DEpendence then
>sh> gives up. But once that's given up, of course, anything goes,
>sh> including that the computational aspects of cognition, having
>sh> already been conceded to be partial, turn out to be minimal or even
>sh> irrelevant...

I hope it is clear by now how cutting computational properties finer may be a way out of the CR. But even if it isn't, C is C is at *least* non-vacuous, even if you think it is false. If the C is C claim is vacuous, how could you take it to be false?

Ron Chrisley

Date: Thu Sep 2 00:00:47 EDT 1993 From harnad@clarity.princeton.edu (Stevan Harnad)

COMPUTATIONALISM: STRONG AND WEAK

In the exchange with Ron Chrisley, a couple of issues have gotten intertwined (partly because it was not just an exchange with Ron Chrisley). The issues that have been under discussion are:

(1) What Is Computation? (2) Cognition Is Computation (True or False?) (3) If Everything is Computation (2) Is Vacuous (T or F?)

I have adopted a provisional answer to (1): (1) Computation is implementation-independent symbol manipulation. This is fairly standard.

Ron seems to be proposing something else -- an "implementation- DEpendent" form of computation -- but this has not been formally worked out as (1) has, so it is not clear what it means.

On a first pass, since the implementation-independence of standard computation was one of its essential features (it might as well have been called "physics-independence" or "differential-equation- independence"), the one that carved out the COMPUTATIONAL LEVEL OF DESCRIPTION, distinguishing it from the PHYSICAL LEVEL OF DESCRIPTION. It would seem that "implementation-DEpendent computation," if it is not an oxymoron, is physics-dependent "computation," a hybrid that can no longer sustain the original motivation for COMPUTATIONALISM about the mind (whose complement, and predecessor, let me remind us all, was PHYSICALISM), namely, to distinguish phenomena that were essentially physical from those (like cognition, on this hypothesis) that were essentially computational.

"Implementation-dependent-computation," in other words, physics- dependent computation, undermines this motivation somewhat, and leaves the anticomputationalist (in the old, implementation-independent sense) unsure whether there is anything left to disagree with.

There's also still the question of truth/falsity vs. vacuity. Let's try some more specific formulations of (3):

"PanVirtualism":

(3') Every physical system is just the implementation of an implementation-independent computational system.

3' is false. (Counterexamples: real furnaces vs. virtual furnaces. The latter lack heat, which is an essential property of furnaces, even though real and virtual furnaces are computationally equivalent.)

Let us suppose, however, counterfactually, that 3’ were true. Consider whether 2’ would then be vacuous:

"Strong Computationalism":

(2’) A cognitive system is just the implementation of an implementation-independent computational system.

Since 2’ would follow from 3’ by universal instantiation alone, and hence would say nothing special about cognitive systems, it would, I suggest, be vacuous.

But since 3’ is false, 2’ is not vacuous. Rather, because of the symbol grounding problem, it is false. (Counterexample: Searle’s implementation of the T2-passing system does not understand -- unless you are willing to believe that some very startling things happen when a person merely memorizes a bunch of meaningless symbols, namely, unbeknownst to him, a second mind is born. Is there any evidence, other than repeating 2’, louder, that such a thing is possible?)

Now here’s an alternative version of 3:

"Physical Church-Turing Thesis":

(3") Every physical system is computationally describable (i.e., computer-simulable, computationally equivalent to some Turing Machine)

This is the physical version of the Church Turing Thesis, and for discrete systems or discrete properties, it is, I think, true. Supposing it is indeed true, consider now whether 2" would be vacuous:

"Weak Computationalism":

(2") A cognitive system is computationally describable.

Since 2" would follow from 3" by universal instantiation alone, and hence would say nothing special about cognitive systems, it would, I suggest, be vacuous. However, it would leave open the question of the actual computational description of cognitive systems. The right description would be true, the wrong one would be false. (This is what Searle would call "Weak AI," and in contrast to [strong] computationalism or Strong AI it should perhaps be called "computabilism" [cf. Harnad 1982, 1989].)

The correct computational description of cognitive systems could, as Ron Chrisley suggested, turn out to be "wildly disjunctive," in which case, though true, it too would be vacuous. Or it could turn out to carve out a natural computational kind (as I believe it would), in which case it would be both true and useful (in understanding and building a cognitive system). But even if the latter were the case, 2" is not 2’, and so there would be virtual cognitive systems that also fit the correct computational descriptions (virtual minds), yet they would not be real cognitive systems, and they would lack mental states.

I think this sorts out the alternatives that have gotten intertwined in this discussion. I will now move to the point-by-point commentary on Ron’s posting, but let me preface it by saying that until it is fleshed out, the nonstandard notion of "implementation-DEpendent computation" can only conflate things again, because it equivocates on just what computation is, and hence just what cognition would or wouldn't be, if THAT were what it was. Nor do I think that wrapping in specific implementational factors such as "timing," or system-external factors such as "environment" or "history" helps clarify matters (all computation is "environment-dependent" to the extent that it is data-driven, but so what?).

I know that the motivation for all this is to preserve the hopeful insight that the "intentional" will somehow be captured by the computational," if not the implementation-INdependent kind, then the implementation-DEpendent kind. But we must not let our hopes get the better of us. Computation is not answerable to cognition. If I'm right, computation's ungroundedness (the symbol grounding problem) blocks its capturing the intentional, and the only remedy may be to hybridize it with physics in the right way (e.g., Harnad 1992, 1993a). There is no need to rechristen that hybrid "ID computation" just to save appearances for a failed hypothesis.

Ron Chrisley (ronc@cogs.susx.ac.uk) wrote:

rc> [E]ven if cognitive computation *doesn't* turn out to be a special rc> kind of computation, that would just mean the "Cognition is rc> Computation" ("C is C") claim would be *false*, not vacuous.

I hope I've shown that it would be vacuous construed as 2' ("Strong Computationalism") if 3' ("PanVirtualism") were true; 3' is false, however, and I have tried to show that 2' is false too. It is is also vacuous expressed as 2" ("Weak Computationalism"), assuming 3" (the "Physical Church-Turing Thesis") is true. A particular computational description of cognition, however, might be (i) vacuous (if "wildly disjunctive"), (ii) false, or (iii) true and useful -- though even in the latter case, not every implementation of it would cognize.

rc> [T]he "implementation-INdependent vs implementation-DEpendent" [I/D] rc> distinction is not meant to be closely related to the "cognitive rc> computation vs non-cognitive computation" [C/N] distinction... I think rc> the non-vacuity of ubiquitous computation holds whether or not the best rc> account of computation is implementation-dependent.

The two distinctions must be closely related, since I/D concerns what computation IS, hence what cognition would or would not be (C/N).

rc> By ID computation, I mean an account of computation that sometimes rc> assigns different computational properties to systems that differ rc> *only* in what is classically (II) thought of as their implementations, rc> and *not* in their traditional computational properties. Thus, ID rc> computational properties (at least sometimes) cut finer than do II rc> properties.

Fine, but what you have not yet justified is calling ID computation COMPUTATION. So far, the view from here is that the finer cutting of ID is in the physics, hence not the computation. (I could redefine "computation" as serotonin release and save C=C that way too, but to what purpose?)

rc> even on an ID account, any two qualitatively identical physical rc> systems will necessarily have the same computational properties, rc> but that would obscure the possibility of ED -- rc> environment-dependent -- computation, which can assign different rc> computational properties to identical systems, as long as their rc> environments are suitably different

We can’t just import philosophers’ favored "wide functionalism" -- which tries to include in the meaning of words/thoughts their causal connections and functional relations with the objects/events/states in the world that they are about -- and dub it a new form of computation. If the meaning of two identical computers’ identical internal states differs as a function of which room they’re in, or what past history brought them to their current identical state, then for me that just underscores the ungroundedness of computation. Widening computation to include the room or the past is just changing the subject.

rc> No, a computer running the sun’s program would *not* have to be very rc> hot. I would only believe that if I thought that *being a sun* was a rc> computational property. But it is not: it is possible that X has a rc> computational description, and that X has property P, and yet P is not rc> a computational property. This would be shown to be the case if there rc> were some systems that also had X’s computational description, yet rc> lacked P.

"Being hot" is surely an essential property of being a sun. What should supervene on running the sun’s (or the mind’s) program depends on whether you are proposing 2’/3’ or 2"/3". What you say above conflates the two.

rc> So a computer simulation of a sun is not (typically) a sun. So *being a rc> sun* would not be a computational property, *even if everything had a rc> computational description*. So it is possible that "being a cognizer is rc> a computational property" is also false, even if everything has a rc> computational description. So computational ubiquity does not render rc> the "cognition is computation" claim vacuous.

Again, conflating 2’/3’ and 2"/3".

rc> ID computation cuts finer than II computation, and says that what II rc> computational theory *thought* to be computationally irrelevant (time, rc> physical constitution, connection with subject matter, etc.) is rc> actually very relevant, and constitutive of certain computational rc> properties.

ID may "cut finer" than II, but is it still computation? What’s relevant to cognitive modelling is not necessarily relevant to what computation is or isn’t. The fathers of the theory of computation (Von Neumann, Turing, Church, Goedel) rightly regarded timing, physical constitution, connection with subject matter, etc. as irrelevant because they were concerned with computation, not cognition. You can certainly REDEFINE "computation" to make it fit cognition, but once you’ve done that, what reason is there to believe that you’re still talking about computation at all?

>
>sh> Indeed, it sounds as if the dependence-story, unlike the
>
>sh> computation-story, will turn out to be mostly physics in that case. > rc> I find that very unlikely, given the intentional nature of cognition. rc> The expression, in the non-intentional language of physics, of a temporal rc> constraint in an ID computational theory (especially of a cognitive rc> system) would be as wildly disjunctive as a physical expression of an rc> II computational constraint.

Timing itself is not a problem for physics (if anything, it’s more of a problem for computation). Making timing depend on content (or vice versa) is no problem either, as long as the content need not be intrinsic (i.e., as long as it can be ungrounded). The trick is in grounding the internal symbols in what they're about, external interpreter-independently. I think altogether too much mystery is being projected here onto things like timing and "history."

rc> [I]f you subscribe to an ID theory of computation, you explicitly rc> *deny* that TM equivalence guarantees computational equivalence! Why rc> should we think that just because an alternative account of computation rc> cuts more finely than traditional computational categories do, that we rc> have left the computational/intentional and landed in the mere rc> physical?

Because, until further notice, the alternative account of computation is a pig in a poke! And cutting more finely just amounts to hybridizing computation with physics -- which I advocate too, I hasten to add, but not with respect to "timing," but with respect to sensorimotor transduction, neural nets, and other analog structures and processes that might help ground symbols in the objects they're about.

To put it another way, your "finer cutting" has to be motivated in terms of COMPUTATION itself, not ad hoc properties -- so far utterly noncomputational (= physical) ones -- that somehow look more hopeful than just "traditional" computation if it happens to be COGNITION (rather than computation) that we're trying to explain. (By the way, even so, I'll bet timing, history and and "environment-dependence" are not even the right noncomputational properties, and that the right ones will turn out to be the ones that generate T3-capacity.) In any case, unless you can make a cognition-independent case for the fact that these "traditionally" noncomputational properties are indeed computational ones, you don't have finer-cutting computation but either circularity or yet another reason to believe that C=C is false.

rc> What is being proposed here (and this is getting more like a defense of rc> ID computation, rather than defending the bite of the C is C claim) is rc> that what II computational theory would take to be two implementations rc> of the same computation are instead implementations of two different rc> computations.

This will need a formal defense, and not in the arena of cognition, but in that of computational theory itself.

rc> Database programs, formal system games, etc. are realized in systems rc> that can be looked at computationally, yet they are not cognitive. The rc> C is C claim does *require* that cognition be realized in systems that rc> can be looked at computationally. But it does not imply that this is rc> what makes cognition special. In fact, it might be that *everything* is rc> realized in systems that can be looked at computationally, yet the C is rc> C claim might be false. Or at least that's what I've argued.

This harks back to the 2'/3' and 2"/3" partition once again.

rc> [In response to the request for] "examples of other rc> phenomena that have been taken to be essentially computational because rc> there turned out to be natural kinds of computational systems": rc> [C]ognition is computation because it can be naturally characterized in rc> computational terms, and perhaps (scientifically) only in those terms. rc> I don't have many other examples, but perhaps something like *sorting* rc> is essentially computational in this sense.

Well, that still doesn't sound like a strong motivation for redoing the foundations of computation. (Also, computers don't sort; they merely manipulate symbols in a way that is systematically interpretable by us as sorting...)

rc> To "suggest what lines the specialness might take in the case of rc> cognition": Of course I can't give THE answer, because cognitive scientists rc> haven't found one (yet?). Well, current orthodoxy says that cognitive rc> systems are those systems that have memory, can learn in these rc> particular ways, can reason common-sensically in these ways, have some rc> type of perception/action link with the world, etc...

This all sounds rather general, and does not yet separate us cognizers from some rather primitive computational systems. And it too seems highly interpretation-dependent.

rc> The hypothesis is that the characterizations that will allow us to rc> distinguish systems that have these essential cognitive features from rc> those that do not will be *computational* characterizations. This might rc> be false; it certainly isn't vacuously true, even if everything has a rc> computational characterization.

It's also not clear what sense of "computation" it refers to.

rc> It might be the case that although rc> everything has a computational characterization, any proposed set of rc> distinctions will fail because either: a) the distinctions are not rc> expressible computationally (e.g. "something must have a soul to be rc> cognitive"); b) the distinctions do not actually divide up the world rc> into the cognitive and non-cognitive (e.g. "something must instantiate rc> Interlisp-D in order to be cognitive"; [probably many cognitive things rc> *don't* instantiate any dialect of LISP!]); or c) both a and b are rc> true. (The examples are deliberately unreasonable and extreme, in order rc> to distinguish the possibilities clearly).

Agreed (as sorted out in the 2'/3' and 2"/3" classification).

rc> ID computation just puts greater restriction on what can count as rc> an implementation of a given computation. So there will still be rc> multiple realizability, of a more restricted sort.

But will it still be just computation?

rc> I've already questioned, above, this assumption of yours, that if a rc> difference isn't a formal computational difference, then it must be rc> merely a physical difference.

But all you've offered by way of support is that there might be ID "computation" still worthy of the name, and that physics alone is not intentional enough. Physics (applied physics, really, engineering) captured what it is to be and make a furnace, car, or plane, although one might have thought those to be too wildly disjunctive. I don't see why reverse engineering (a robotic hybrid of physics and computation) can't capture T3 capacity too -- and I think a mind will piggy-back on the successful outcome (Harnad 1994).

>
\>sh> "Not wildly disjunctive" just isn't a positive enough
>
\>sh> characterization to give me an inkling. Do you have examples, or
>

>sh> (relevant) analogies? > rc> Imagine two systems that are formally computationally identical, yet rc> their environments differ. Often, if these systems are sufficiently rc> interactive with their environments (e.g. feedback-oriented), we will rc> find it useful to classify them into different rc> informational/computational states. What would have to be identical rc> computational systems from an II point of view can be seen, from an ID rc> (or ED [environment-dependent], see above) point of view, distinct.

This is already true of two "smart" servomechanisms in two different environments (perhaps even two computers running the same program on locally identical but globally distinct data sets). It does not necessitate new forms of computation (ID), nor does it capture cognition. It just raises a wider (extrinsic) interpretation question than a narrow, stand-alone computer account can answer. So much the worse for pure (ungrounded) computation, I would say...

rc> ubiquity of computation is not a problem for the C is C claim even on rc> an ID view of computation. On such a view, there can be systems that rc> are II computational (TM) equivalent, that are not cognitively rc> equivalent. But ID theorists could still hold that *ID* equivalent rc> systems must be cognitively equivalent.

But all of this only at the cost of redefining computation in the service of cognition -- which hardly clarifies cognition (or computation).

rc> I’m not sure if the ubiquity of computation is the same as the rc> Church-Turing thesis. But if you deny the ubiquity of computation (deny rc> that everything has a computational description), then you can *ignore* rc> my claim! Because my claim was conditional: Even *IF* rc> everything has a computational description, the C is C claim is not rc> vacuous.

I don’t deny it -- I accept it (it’s 3") -- yet it still leaves "C = C" (construed as either 2’ or 2") as vacuous (and "C = ID-C" as ad hoc and uninformative).

rc> (To address the Chinese Room argument very quickly: I agree that this rc> spelling out of ID computation should be done. But as long as it is rc> even a mere possibility, then Searle’s argument fails. Searle is the rc> one making the strong claims; he has not *shown* that these claims are rc> true until he shows us that there is no possibility of ID computation.)

Not at all. His argument was explicitly about II computation. He put it quaintly: "Strong AI" amounted to three propositions (each is followed by a translation into what he really meant, in the terms of this discussion):

(a) "The mind is a computer program" (i.e., cognition is computation) (b) "The brain is irrelevant" (i.e., computation is implementation-independent) (c) "The Turing Test is Decisive" (i.e., every implementation of a T2-passing symbol system understands)

The Chinese Room Argument is then based on Searle’s BECOMING an implementation of the Chinese T2 passing symbol system (by memorizing and executing all the symbol manipulations) without understanding Chinese (with no other plausible candidate to impute any understanding to). The argument would be nonsense if computation were not implementation-independent (i.e., not-b) -- but then Strong AI (Strong Computationalism) would be false anyway, and that’s all Searle was trying to argue.

So there is definitely no question of Searle’s having the burden of trying to show anything about some inchoate ID "computation" that no one has yet championed or specified (and that may not amount to computation at all).

rc> On the ID view, one *does* only have to look at what is computationally rc> relevant in order to determine cognitive properties; it is just that rc> the ID theorist considers *more* distinctions to be computationally rc> relevant than does the II theorist. It could be that these extra rc> distinctions *are* non-computational, non-intentional, merely physical. rc> That would imply that the ID view of computation is incorrect. But we rc> don’t know that it is incorrect right now. And even if it is incorrect, rc> the ubiquity of computation does not make the C is C claim vacuous!

We not only don’t know whether the ID view of computation is correct of cognition (or of computation) -- we don’t even know what it is. But this all makes it sound as if computation were somehow answerable to cognition, and that seems to be getting it backwards...

And if ID computation is indeed just an ad hoc hybrid of the computational and the physical -- i.e., not a form of computation at all -- then that DOES leave C = C vacuous, whether one subscribes to 3’ (PanVirtualism) or 3" (the Physical Church-Turing Thesis).

rc> I shouldn’t have mentioned ID computation at all, it seems! I guess the rc> only reason to mention it is if: [i] you think computation is rc> ubiquitous; [ii] you think a Hayes-style response is the only way to rc> respond to the Chinese Room; [iii] you want the C is C claim to be rc> non-vacuous. Since you, Stevan, don’t meet at least 2 of these rc> conditions, it is a bit irrelevant.

No, I think I meet all three of these conditions: (i) I subscribe to the Physical Church-Turing Thesis (3"). (ii) I think a Hayes-style response (to the effect that Searle’s is not a valid form of implementation) would be the only way to respond to the Chinese Room (but I think this response, like ID, is moot until Hayes comes up with a nonarbitrary criterion for what counts as an implementation). And (iii) I think the C is C claim is nonvacuous in the form of Strong Computationalism (2’), given that PanVirtualism (3’) is false; unfortunately, though nonvacuous, 2’ too is false.

So I think your own invocation of ID is not only relevant, but necessary (for you), but also that it’s as moot as Hayes’s strictures on implementation -- and for roughly the same reason: You’re not allowed to invoke arbitrary cognition-dependent constraints on computation or implementation JUST in the service of saving C = C!

rc> I meant stronger in the logical sense. An ID notion of computation will rc> mean that even fewer systems are cognizers, which is even further away rc> from the panpsychism that C is C, plus computational ubiquity, rc> supposedly entails.

Only PanVirtualism (3’) entails panpsychism. The Physical Church-Turing Thesis (3") merely entails that Weak Computationalism (2") is trivially true (i.e., vacuous). And ID "computation" does indeed mean that fewer systems are cognizers than implied by Strong Computationalism (2’), in fact, fewer systems than are (II) COMPUTATIONAL ones, which would thereby refute Strong Computationalism -- if only we knew what ID computation WAS (and why it is called "computation").

>
>sh> A TM performing implementation-DEpendent computations. I
>
>sh> can't imagine what to imagine! > rc> Well, it's an instantiated TM, but it may have a different environment, rc> etiology, and/or temporal character than you do. If so, it may be rc> computationally distinct from you, even though it is TM-equivalent.

I'm afraid that doesn't help. It just seems to be forcing the concept of computation, for some unfathomable reason, into the mold of cognition (or rather, HYPOTHESES about cognition)! Why on earth would one want to call two TM-equivalent TMs COMPUTATIONALLY distinct on the grounds of such TM-external factors as where they're located or what their history was (if their history left them TM-equivalent)? (Actually, I think STRONG equivalence would be better to use here than TM-equivalence, which is only I/O equivalence.) Ditto for timing factors: They're differences all right, but why COMPUTATIONAL differences?

N.B.! Remember that in Searle's Argument he does not just produce the same I/O as the T2-passing program, he actually EXECUTES the same program, symbol for symbol, step for step. This, I believe, is called strong or internal equivalence. It is not to be confused with I/O or weak or black-box equivalence (which is all that T2 or T3 draw upon). (Please reflect a while on this point; it's a subtle and critical one in all this, and has not been made explicitly enough yet. Turing Indistinguishability should not be confused with computational [Turing] equivalence, and especially not strong equivalence.)

rc> An ID computational theory could lump together as computationally rc> equivalent systems that are T3-distinguishable.

And systems that are strongly equivalent. Not a desirable outcome, I think:

First, "T3-distinguishable" is equivocal. Since we normally speak of T3-INdistinguishability (i.e., a candidate we cannot distinguish from a real person on the basis of either his symbolic or his robotic performance for a lifetime): Do you mean a candidate we CAN so distinguish? But then he's failed T3! You probably just mean T3-irrelevant T3-performance differences: the kinds of things that DIFFERENTIATE me from you (or a T3 robot) but that do NOT allow any of the three of us to be singled out as the one without the mind.

Well, if, charitably, we assume you mean the latter, what you should have said was T3-indistinguishable T3-performance differences might be lumped together as equivalent by "ID computational theory." Well, if so, that's good, but they will already have been lumped together in virtue of being T3-indistinguishable! (And, ceterum sentio, we have yet to hear the non-ad-hoc principles of "ID computational theory"...)

The other part of the outcome is not desirable either: If ID computation distinguishes what is II equivalent, that's bad prima facie news for the prospect of the differences being computational ones. It's even worse if it distinguishes what is strongly equivalent.

rc> all you have to imagine is a theory of computation that cuts finer than rc> formal accounts do nowadays.

Call it a failure of imagination, but I have to SEE the theory, and the explanation of why it's a COMPUTATIONAL theory.

>
>sh> (Again: multiple-realizability is not the same as
>
>sh> implementation-independence.) I have no problem with synthetic heat,
>
>sh> but that still does not help me see heat as computation (the real
>
>sh> problem is VIRTUAL heat). And even if there is a nice, crisp
>
>sh> ("nondisjunctive") set of unique computational invariants that
>
>sh> characterize hot things and no others, I still don't see what it
>
>sh> would mean to say that heat was computation -- except if EVERY
>
>sh> implementation of the heat program were hot -- which is decidedly
>
>sh> not true (because virtual heat is not hot). > rc> But every implementation of the heat program *would* be hot, if we rc> grant your hypothesis that there is a computational characterization rc> that captures hot things and no others.

Again, this apparently miraculous outcome on the basis of mere hypothesis is based on the equivocation running through all of this, concerning whether we mean computation in the sense of 3' (PanVirtualism) or 3" (the Physical Church-Turing Thesis), and II or "ID." It's simple to show this: First get 3' out of the way (with the counterexamples above), then adopt 3" instead, and you will always have at least one strongly equivalent "doppelganger" (a virtual, purely computational one) for real properties like heat, one that will lack heat's essential property even if computational properties nondisjunctively distinguish all the real cases of heat from the real cases of non-heat (in the sense of 3"): Real and virtual heat will be strongly equivalent (and computationally distinct from real and virtual non-heat), but only one of them will be hot.

>
>sh> [J]ust as I have indicated that I of course realize that "all cog
>
>sh> is comp" does not imply "all comp is cog," I don't think that the
>
>sh> emphasis on the implementation-independence of cognition was enough
>
>sh> to uniquely characterize cognition, for there are of course plenty
>
>sh> of other implementation-independent computational phenomena. It had
>
>sh> more of a necessary- than a sufficient-condition flavor -- though of
>
>sh> course necessity was not at issue at all. > rc> Then II is neither sufficient NOR necessary for cognition?

This is putting it more formalistically than necessary. Pylyshyn and the other strong computationalists did not think that cognition had to be the ONLY computational natural kind in nature. They just thought that it shared, with whatever other phenomena were essentially computational, the fact that they were implementation-independent: that their essential properties were their computational ones and not their physical ones, and that any physical system that had their computational properties would have their essential properties.

This certainly does not make implementation-independence SUFFICIENT for being cognitive (since other phenomena might be implementation- independent too). The reason I said it was more like a necessary condition, though its necessity was not the issue, was that the strong computationalists were in fact looking for the RIGHT theory of cognition. THAT would be where the SUBSTANTIVE necessary conditions for cognition resided, not in the general computationality stricture. In this respect, their doctrine and even their actual research converged to a degree with that of the weak computationalists.

Where weak and strong computationalists diverged was in what they would claim of a virtual mind -- a computer program that could pass either T2 or a VIRTUAL-WORLD version of T3 (which, of course, is no T3 at all). Strong computationlists would say that the virtual system really had mental states (because cognition is just a kind of implementation-independent computation), whereas weak computationalists would not. (As I've said before, I think the weak computationalists are right, because of the symbol grounding problem and Searle's argument, and that the strong computationalists are just seduced by the fact that thought, unlike heat, is unobservable.) Of course, this difference dictates big differences in methodology too (e..g., I doubt that virtual worlds are the best places to do T3 robotics, though some of it can no doubt be done there; also, a grounded bottom-up approach is not likely to keep running into the frame problem; Harnad 1993b).

>
>sh> What Pylyshyn was saying was that mental states were unlike other
>
>sh> kinds of physical states, and were like other kinds of computational
>
>sh> states (including nonmental ones) in being IMPLEMENTATION-
>
>sh> INDEPENDENT. That wasn't SUFFICIENT to make every computation
>
>sh> mental, but it was sufficient to distance cognitive science from
>
>sh> certain other forms of physicalism, in particular, the kind that
>
>sh> looked for a hardware-level explanation of the mind. > rc> But if you take Searle seriously, you must think Pylyshyn is wrong. rc> So you should leave open the possibility of ID computation: a rc> computation that cuts finer than does II computation (or use the rc> Systems Reply -- oops!).

I do think strong computationalism is wrong, but I don't think ID computationalism is a coherent alternative to it (and I think the System Reply is at best a symptom of misunderstanding Searle's argument, at worst a symptom of the hermeneutic fantasy that the unobservability of mental states in other systems has left us susceptible to).

rc> [I]n Chapter 9 of _The Rediscovery Of The Mind_ (p 209), Searle admits rc> that there are people (e.g. Brian C. Smith of Xerox PARC) who are rc> trying to come up with more embodied notions of computation that *do* rc> place non-formal constraints on the "implementation" when taxonomizing, rc> and that such notions of computation escape his vacuity argument. If anyone comes up with a principled COMPUTATIONAL alternative to implementation-independent computation, that form of computation will be immune to Searle's Chinese Room argument, which depends ESSENTIALLY on the implementation-independence of computation (see Harnad 1993a). As far as I know, though, Brian Smith has not yet come up with this alternative, just as Pat Hayes has not yet come up with the principled definition of "implementation" according to which Searle's is not a legitimate implementation.

On the other hand, I think Searle's suggestion that C=C is vacuous is wrong -- if we construe computation in the standard (II) way and simply reject PanVirtualism (3') as false, and construe C=C as Strong Computationalism (2'), which is then simply false, because of the symbol grounding problem and Searle's Argument (and even if the Physical Church-Turing Thesis, 3", and hence also Weak Computationalism, 2", the latter indeed vacuous, are true, and whether or not 2" is fruitful).

Besides, Searle's Chinese Room Argument would be supererogatory if Computationalism were vacuous.

>

\>sh> Now that you've disavowed UTMs, pure formal syntax and

>

\>sh> implementation-independence, this does not bother you,

>

\>sh> Ron; but just what, exactly, does it leave you with? > rc> It leaves me with the possibility of a computer science that takes rc> seriously the environment, time, physical realization, and -- who knows rc> -- perhaps history. Which is just as well, since that is the kind of rc> computer science that some CS researchers are putting forward as rc> necessary to explain today's complex information processing systems.

So far, it sounds analogous to a new kind of physics in which, instead of idealized invariant laws of nature, we would have only boundary-condition-dependent constraints. A possible outcome, of course, in both cases, but it would have to have strong, principled motivation, in both cases, and demonstrable empirical and theoretical power.

rc> I think the primary dispute here has become: "how could one have an rc> implementation-dependent notion of computation, given that rc> computational properties are *defined* to be implementation- rc> independent?" I reject that computational notions are defined to be rc> II. To say so is to confuse the map with the territory. In discussing rc> what computation is, my principal allegiance is to the phenomenon of rc> computation, what goes on in IBM's, token rings, etc. If the best rc> non-physical, intentional, representational account of that phenomenon rc> individuates properties differently than does a formal account, then rc> that just shows that computation isn't just formal symbol manipulation.

But DOES the best account do that?

rc> When I consider the C is C claim, I don’t take it to be the claim that rc> "Cognition is formal symbol manipulation, which may or may not be the rc> best account of computational phenomena" but rather "Cognition is rc> computation, and the best account of computation may or may not be in rc> terms of formal symbol manipulation". That’s why a notion of ID rc> computation makes sense.

Fair enough, but so far I think the best account of computation is still the formal theory of computation (which is II). There are certainly many legitimate hardware considerations in computer science, but do any of them dictate a reformulation of the formal theory of computation?

rc> *IF* C is C, then the correct analysis of the brain will be at a level rc> higher than the mere neuro-molecular. Just as an engineer can look at rc> bits only, and not see the computational pattern that explains an IBM, rc> so too Searle can insist on looking at neuro-molecular stuff and ignore rc> the computational patterns in the brain.

Neuromolecular analysis is T4. Searle seems to have thought that that was the only alternative to T2 and computation. But the T3 and grounded hybrid system alternative is still wide open too. T4 could be the wrong level even if computation is not the right one.

>

>sh> My own guess is that going "embedded" or "implementation-dependent"

>

>sh> amounts to conceding that the cognition is in the physics rather

>

>sh> than just the computation. > rc> Your *guess* might be right. But so might the *guess* that these rc> implementational factors are themselves computational/intentional, and rc> are not best understood with a mere physical vocabulary.

Physics includes engineering (as in cars, planes, furnaces, transducers, servomechanisms, etc.) as well as reverse engineering (as in T3 robotics). "Computational" certainly does not equal "intentional," but either way, what we need is not a "vocabulary," but a way to understand and perhaps build systems whose "aboutness" is not just projected on to them by us, but grounded, directly and external-interpreter-independently, in what it is that they are about.

rc> Surely there is little that a Martian and I need have *physically* in rc> common for us to both be cognizers. It seems that the best level of rc> analysis of our similarities would be above the physical, but below the rc> formal (if the Chinese Room bothers you, for instance).

But I think the only "level" below the formal is the physical. Planes are functionally equivalent in a PHYSICAL sense that does not leave us attributing flying to molecules. I think something like that should be true of cognizing as well (though, in focusing on SYMBOL grounding, I am actual committing myself to a formal-physical hybrid).

>

>sh> To resolve the ambiguity inherent in computationalism it is

>

>sh> therefore not enough to point out, as Ron has done, that the

>

>sh> "Cognition is Computation" thesis just claims that "Cognition is a

>

>sh> KIND of Computation"; for what it really claims is that "Cognition
>
>sh> is JUST a Kind of Computation."
>
>sh> And it's that "JUST" that I think implementation-DEpendence then
>
>sh> gives up. But once that's given up, of course, anything goes,
>
>sh> including that the computational aspects of cognition, having
>
>sh> already been conceded to be partial, turn out to be minimal or even
>
>sh> irrelevant... > rc> I hope it is clear by now how cutting computational properties finer rc> may be a way out of the CR.

Alas, it's not only unclear how cutting computational properties finer may be a way out of the Chinese Room, it isn't even clear why it would be a way of computing. I'm ready to concede that the "hybrid" system that passes the T3 may turn out to be no more of a computational system than a plane is -- but you should be prepared to concede the same of your ID computation.

Stevan Harnad

Harnad, S. (1982) Neoconstructivism: A unifying theme for the cognitive sciences. In: Language, mind and brain (T. Simon & R. Scholes, eds.) Hillsdale NJ: Erlbaum, 1 - 11.

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Harnad, S. (1993a) Grounding Symbols in the Analog World with Neural Nets. Think (Special Issue on Machine Learning) (in press)

Harnad, S. (1993b) Problems, Problems: The Frame Problem as a Symptom of the Symbol Grounding Problem. PSYCOLOQUY 4(34) frame-problem.11.

Harnad, S, (1994) Does the Mind Piggy-Back on Robotic and Symbolic Capacity? To appear in: H. Morowitz (ed.) "The Mind, the Brain, and Complex Adaptive Systems.

Date: Sun, 24 Oct 93 15:24:11 EDT From: "Stevan Harnad" To: brings@rpi.edu Subject: Symbol Grounding Discussion: Is English a Formal System?

To: Symbol Grounding List

This discussion about whether or not natural languages are (among other things) formal symbol systems has gone a couple of iterations in camera with Selmer Bringsjord (sb), but I (sh)think it's time to open it to the Group. Everything below should be self-contained. Selmer thinks that there are properties of formal systems and properties of English that make it clear that English cannot be, all or in part, a formal system. I think it’s obvious that English is IN PART a formal system (first, because it artificial languages are just parts of English, and second, because English can be de-interpreted, and then all you have left is the formal symbols), but it’s also obvious (because of the symbol grounding problem) that it is not JUST a formal symbol system, because purely formal symbol systems are ungrounded, and English sentences, when we THINK them in our heads as opposed to their simply appearing on paper, are grounded. Please feel free to add to the discussion. -- Stevan

>sb> You claim (H!) that English is a formal system... This is a view I find
>sb> enormously interesting; I hadn’t heard it before. I know, of course,
>sb> because I know your style, that... the view fits you like a glove. Now
>sb> you keep wanting to say, "Well, but I’m not just saying that English is
>sb> a formal system, I’m saying that and ..." For purposes of this
>sb> dialectic, I don’t much care about the "..." -- because I know that if
>sb> P is false then P & Q is false.

You have a scope problem here for "just": Compare your "I’m not JUST saying that English is a formal system" and my "I’m not saying that English is JUST a formal system." Your P & Q conjunction is only relevant for the construal with the wrong scope. Compare: "Selmer weighs 50 pounds" (P’) and "Selmer weighs JUST 50 pounds" (P). The first is true, the second false. My statement about English being a formal system but not JUST a formal system, but rather a GROUNDED formal system, is like the negation of P (or the affirmation of P’ AND Q [groundedness], if you like). There is no point trying to refute it with arguments against P’ construed as P!

>sb> All this grounded stuff is stuff I’m trying to keep out of this particular
>sb> discussion (b/c at bottom you’re just looking to define a causal
>sb> relation R holding between symbols in some symbol system, the agent
>sb> or robot involved, and the external world, and you think that
>sb> such a relation is going to solve the fundamental problem of
>sb> intentionality, of mentality -- but this was, in my opinion refuted
>sb> by Roderick Chisholm years ago).

Please re-read my reply to Searle and others in our joint Symposium on Connectionism vs. Symbolism (Harnad 1993, Searle 1993). I keep repeating (but you, alas, keep not hearing) that GROUNDEDNESS DOES NOT EQUAL OR GUARANTEE MEANING. It is neither a sufficient condition for meaning nor a demonstrably necessary one (only a probably necessary one, by my lights).

Groundedness is just groundedness. It answers only one objection: the objection that in a purely formal (i.e. ungrounded) symbol system the connection between the symbols and what the symbols are interpretable as being about is ONLY in the mind of the interpreter. With the T3-scale robot, this connection is causal and direct, hence the symbols are grounded. However, there may still be nobody home in there, hence no real meaning. (You repeatedly make the same misconstrual of T3: It’s not that a T3-passer MUST have a mind; no sufficient or necessary conditions; just that YOU have no better or worse grounds for denying [or affirming] it has a mind in the case of a T3 robot than in the case of your wife Elizabeth or me.)

>sb> Now P is false -- ’English is a formal system’ is false -- because it
>sb> fails to have certain conditions necessary for something X to qualify
>sb> as a formal system. One of these conditions is that formulas in
>sb> X must admit of unabmbiguous truth or falsity under any
>sb> fixed interpretation. But I’ve proved by counter-example (following
>sb> the common refrain in this regard about world knowledge) that
>sb> English doesn’t have this property:

I don’t think your counterexample (of the semantically ambiguous sentence "Harnad likes flying planes") proves anything. The purely formal subsets of English (e.g., propositional calculus) do have the property of unambiguous truth or falsity inder a fixed interpretation, and so does the rest of English -- if you first DE-INTERPRET the well-formed formulas, reduce them to squiggles and squoggles. Consider this analogous to paring it down to its core 50 pounds. That’s not ALL that English is, but it IS that. (The existence of polysemy just indicates that the notational system of natural languages is a bit hasty and takes some short-cuts -- creating ambiguities that could easily be circumvented with a few trivial extensions of the lexicon: So what?)

Your "counterexample" rides on semantic ambiguity. I don’t know what semantics is, but the formal symbols are supposed to be deinterpreted. Without the interpretation, "Harnad likes flying planes" becomes P, or HLFP. So what then?

>
>sh> Please specify what it is that I am affirming or denying, by
>
>sh> your lights, when I say that "The cat is on the mat" (1) is a sentence
>
>sh> in English, (2) is interpretable as meaning that the cat is on the mat,
>
>sh> (3) is a string of formal symbols that is interpretable as claiming that
>
>sh> the cat is on the mat, (4) is either true or false (but would not be if
>
>sh> it instead said "This sentence is false" -- though it would still be
>
>sh> English, and still a string of symbols), (5) and is not ONLY a string
>
>sh> of formal symbols interpretable as claiming that the cat is on the mat
>
>sh> when, instead of being a string of formal symbols on paper, it is a
>
>sh> thought, thought in my head. >
>sb> You pulling my leg? Take them in turn. My daughter knows (1); she’s
>sb> 6 (and not a prodigy w/ command over the terrain of logical systems).
>sb> Ergo (1) isn’t related to H!.

I don’t get the point here. Is everything I say supposed to disagree with whatever your daughter knows?

>sb> (2) is affirmed by any proponent of the correspondence theory of truth,
>sb> and such thinkers often don’t know more than my daughter does about
>sb> logical systems.

Ditto.

>sb> (3): Now we start to kick up some substance. I suspect you would
>sb> defend (3) solely on the basis that English has a fixed alphabet
>sb> E, and that the sentence in question is a string over E. Problem is,
>sb> here we go with the Harnadian informalist take on things. This
>sb> rationale carries no weight in the context of formal languages.
>sb> After all, what’s the language? Specify it for me, right now.
>sb> Is it context-free? Context-sensitive? Between them? You see,
>sb> when you use ’formal symbols’ you trade on a cozy little equivocation,
>sb> that between the nice mathematical sense, and the naive, intuitive
>sb> sense (on which (3) is obviously true). So I dispute the formal
>sb> sense of (3).

I don’t know what context free/sensitive has to do with it. A formal system is a system of arbitrarily shaped squiggles and squoggles that are systematically manipulable on the basis of formal (shape-governed) rules [algorithms] operating on their arbitrary shapes in such a way as to make the symbols and symbol strings (and perhaps even the rules and manipulations) systematically interpretable -- i.e., construable as being ABOUT something (e.g., numbers, truth, cats, mats). That’s all I’m saying when I say ("H!") English is (in part) a formal system (now YOU’re the one talking about formal "languages" instead of just formal "systems," but I don’t care). I don’t inherit consistency, completeness, halting or decidability problems in saying all this, except inasmuch as I am saying they too are a PART of English. So what?

>sb> (4): I won’t quibble here. Again, the man on the street will
>sb> gleefully affirm this. Has little to do with H!.

I have no idea what you’re reading into "H!" but I just got done saying what I mean by it, and if your daughter and the man in the street agree with me, that’s just fine with me.

>sb> (5): Don’t for the life of my understand this one. Come again?

"The cat is on the mat" is not ONLY a string of formal symbols interpretable as claiming that the cat is on the mat when, instead of being a string of formal symbols on paper, it is a thought, thought in my head.

Short version: English is not JUST a formal system (and you don’t weigh JUST 50 pounds).

>sb> If I could accept the claim that a formal system like first-order
>sb> logic was a proper subset of English, I could welcome this.
>sb> But you go wrong here, surely. Don’t be confused by the fact
>sb> that the metalanguage used in mathematics to reason about first-order
>sb> logic contains English expressions. That doesn’t in the least
>sb> imply that the metalanguage is English! Who, but you -- formidable
>sb> you, I concede -- thinks *this* is true!?

Here are some English sentences for you: "The cat is on the mat." "2 + 2 = 4." "Ax(Mx --> mx)". "‘The cat is on the mat’ means the cat is on the mat." "‘2 + 2 = 4’ means that two and two are four, that twice any quantity is four times that quantity, etc." "‘Ax(Mx --> mx)’ means that for all x if x is a Man then x is mortal," etc.

ALL of the above are a just bunch of symbol strings, ALL of them are in English, including the mathematical and logical ones, and both the object-level and meta-level ones. They are also all UNGROUNDED strings, sitting there on the screen or a piece of paper. When they are instantiated as thoughts in my head, however (i.e., the thought that the cat is on the mat, say, or that twice two is four, or that all men are mortal), the stuff that is actually going on between my ears in real time at the time that thought is being instantiated is not JUST the instantiation of the symbol string. It is the instantiation of a grounded symbol string.

I have no idea what is actually going on in my head when that thought is being thought. My hypothesis (worked out a bit more concretely in my hybrid model for category learning) is that what makes that thought be a thought about whatever it is about is certain other structures and processes (over and above the symbol tokens themselves) that are co-activated during that thought, and further structures and processes causally linked to those, actually or potentially (in my hybrid model, these include analog sensorimotor projections and invariance-learning neural nets). Those structures and processes correspond to the substrate of my T3 capacity.

>
\>sh> I don’t know what baggage attaches to the "thesis" in logic and
>
\>sh> philosophy, but for me it’s no more than that natural language really
>
\>sh> IS a string of (systematically intepretable) formal symbols, but not
>
\>sh> JUST that; it’s also grounded -- and it’s there, in the grounding, and
>
\>sh> not in the formal, syntactic properties, that the real cognitive
>
\>sh> "action" is. >
\>sb> I know where you stand. You accept a good, solid conception of
\>sb> consciousness and the mental -- but you pin all your hope on
\>sb> the causal relation I mentioned above. Ever try to specify the
\>sb> relation? Ever try to characterize it with *some* precision?

I’m trying to do what I can, in my empirical modelling. What do philosophers do?

\>sb> Brentano said that intentionality was the foolproof mark of the
\>sb> mental; if a thing could genuinely have beliefs about something,
\>sb> it was mentally alive, conscious. He also said, as many others
\>sb> have since, that there is no causal relation, no matter how
\>sb> complex, that makes my beliefs about you *bona fide* someone’s-home
\>sb> beliefs about you. You will find, you think, this relation
\>sb> holding between logical systems, the physical world, sensors,
\>sb> effectors, etc. It’s all blind faith -- faith attacked a century
\>sb> ago.

Brentano and Wittgenstein #2 and others have SAID such things, I know, but have they ever taken on the empirical problem of actually GENERATING the requisite causal connection? I certainly don't see anything like a PROOF that they cannot be generated. Besides, there is an equivocation on WHAT it is our responsibility to generate. I aim at T3. That's a real problem, one that has clearly been solved by evolution, and, once we accomplish the requisite reverse engineering, it's solution will tell us as much as it will be possible to know about the mechanics of the mind. If it is not enough (as it may conceivably fail to be), I rest my case, for I hold that we cannot hope to be any the wiser. But I certainly don't take arguments against the logical sufficiency of T3 to be arguments against pursuing T3! (What, for example, is the empirical alternative?)

>sb> So, I think H! is provably false, though it's nice that you affirm
>sb> your list of (1)-(5). And, as you can see, you keep bringing
>sb> everything back to your grounding stuff, which I think is just
>sb> plain false, for old, old reasons, just as devastating now as ever.

If you are more careful about what "H!" is and is not claiming, you may lose your assurance that it is false. Remember, reverse bioengineering and philosophy are not the same thing, even if they have some overlapping interests.

>sb> Take care of yourself. Tell France no. The Northeast needs you
>sb> here physically. (Doesn't BBS, anyway?)

In the virtual world of the Internet, I am everywhere at once. It makes no difference, really, where my physical body resides.

Chrs, Stevan

Harnad, S. (1993) Grounding Symbols in the Analog World with Neural Nets. Think 2: 12 - 78 (Special Issue on "Connectionism versus Symbolism" D.M.W. Powers & P.A. Flach, eds.).

Searle, J.R. (1993) The Failures of Computationalism. Think 2: 12 - 78 (Special Issue on "Connectionism versus Symbolism" D.M.W. Powers & P.A. Flach, eds.). Pp. 68-73.

Date: Tue, 26 Oct 93 20:40:03 EDT From: "Stevan Harnad" To: phayes@cs.uiuc.edu Subject: Re: "Is Computation Cognition?"

Here is a belated reply of Hayes to Bringsjord, its posting retarded by my transcontinental peregrinations. Apologies. SH.

Date: Mon, 6 Sep 1993 19:17:44 +0000 From: phayes@cs.uiuc.edu (Pat Hayes)

In response to Selmer:

>
>ph>... I take it that you and I are, in fact, >
>ph> computing machines (as well as other kinds of machine). >

>sb> That’s an interesting move. I shall have to remember to advise my
>sb> religious friends to make a parallel move when debating atheists:
>sb> Begin by assuming that God exists. All of "ambitious" AI/Cog Sci
>sb> hinges on the claim that human persons are computing machines.

That was my point: these arguments are worthless because they amount to little more than the observation that our positions are inconsistent with one another. If I stick to my assumptions then they show yours are false, and vice versa. We knew that already.

>
>ph> This is an example of a class of anti-AI arguments of the general >
>ph> form ’People have mysterious property X. Automata don’t have X. >
>ph> Therefore people aren’t automata." >
>sb> Where, pray tell, did I say the property in question (being such that
>sb> one can experience the raw feel of meeting a long lost friend, e.g.) is
>sb> mysterious?

You didn’t. I did. I find no meaning in the term ’raw feel’.

>sb> In WRC&CB I point out that such properties undergird a
>sb> normal appreciation of creative text. Your position would then be
>sb> that appreciating a short story is mysterious. I won’t even complete
>sb> the *reductio*.

Thats not reductio: indeed, appreciating a short story IS a mysterious act. Do you have a complete theory of story-appreciation available?

>sb> I know the literature on introspection well. I know all the attacks on
>sb> introspection from case histories a la Sacks, insanity, weakness of the
>sb> will, drug use, RT studies galore in cog psych, David Hume, Paul
>sb> Churchland spy-behind-enemy-line cases, and on and on *ad
>sb> nauseum*. In *What Robots Can & Can’t Be* I defend the proposition
>sb> that human persons enjoy what I call ’hyper-weak incorrigibilism.’

Ah. OK, I misunderstood your position, it wasn’t quite what you stated it to be. No matter, this falls under my case number one: a strange property which people have. Back to modus tollens.

>
>ph> Selmer also gives us an outline of his argument from free will. >
>ph> This uses the concept of ’iterative agent causation’, and concludes >
>ph> that this is true from a premise concerned with moral >
>ph> responsibility. Even if we assume the truth of it, however, his >
>ph> conclusion doesn’t follow. The proposition is that people, not >
>ph> events, bring about certain events, and bring about their bringing >
>ph> about of those events. Lets accept that, whatever it means. He >
>ph> concludes that people aren’t automata. Why? Surely automata >
>ph> might also bring about certain events, and bring about their >
>ph> bringing about of those events. >
>sb> And bring about the bringing about of the bringing about of those

>sb> events, *ad infinitum*? Did you catch the three dots?

Ah, no I didn't. But that's OK: I'm quite willing to accept that if any physical entity can do this, then an automaton can. The reason for the qualification is that I think I now see your strategy, since this gives us an infinite number of states the person has to be in. If this is supposed to correspond to an infinite number of physical states, then I will argue that we can't do it either, since nothing can do it. If on the other hand these mental states don't correspond to physical states, then you seem to be arguing for dualism. Or maybe this notion of 'mental state' is just kind of, well, vacuous?

>sb> Of course, what's in the background is the argument of which this
>sb> was an *outline*. The states in question correspond to bringing
>sb> about certain mental events. The idea is that we are forced to
>sb> countenance something which looks massively counterintuitive (that
>sb> persons enter an infinite number of mental states in a finite amount
>sb> of time). Turing machines allowed to do that could, e.g., solve the
>sb> busy beaver function.

It depends what you mean by a 'mental state'. If a mental state is a computational state, yes. But since people apparently perform this little miracle every time they decide to buy a pizza, I'm quite happy to say that my Macintosh does it every time it decides to switch to AfterDark. Of course this infinite stack of mental state doesn't correspond to any computational state, but that shouldn't bother you.

(Incidentally, computational ideas suggest an alternative way around this problem. Maybe the mental state should be characterised as recursive rather than infinite: that is, it brings itself about as well as bringing it about that... To a thinker familiar with mathematical ideas this suggests infinity, but Cognitivist insight suggests something more tractable. And then your argument loses its force, of course, since we can put automata into recursive states.)

>
>ph> I might add that the premise "were they >
>ph> automata" here is unnecessary, since if people can enter an >
>ph> infinite number of states in finite time then they could perform >
>ph> physically impossible feats no matter what they were, which >
>ph> suggests - if it needed suggesting - that they probably cannot do >
>ph> this. >
>sb> Iterative agent causation has only been maintained by thinkers who
>sb> see no other way to make room for freedom (e.g., Roderick Chisholm,
>sb> Richard Taylor). Since I know, if I know anything, that whether or
>sb> not I have pizza tonight is up to me, and I also know that
>sb> determinism, indeterminism and normal agent causation imply that
>sb> whether or not I have pizza tonight *isn't* up to me, something like
>sb> iterative agent causation comes on stage.

So, if Cognitivism is correct then Selmer Bringsjord doesn't KNOW that he is ABSOLUTELY FREE to decide to buy pizza. Seems reasonable to me.

Pat Hayes, Beckman Institute. University of Illinois, Urbana

CORRELATIONS ARE JUST THE CHESHIRE CAT’S GRIN Comments on Pindor’s Comments on the Symbol Grounding Problem

Stevan Harnad Cognitive Science Laboratory Princeton University 221 Nassau Street Princeton NJ 08544-2093 harnad@princeton.edu

> COMMENTS ON THE SYMBOL GROUNDING PROBLEM
> gopher://gate.oxy.edu
> /0ftp%3amrcnext.cso.uiuc.edu%40/etext/ippe/preprints/Phil_of_Mind
> /Pindor.Comments_on_the_Symbol_Grounding_Problem/symbol.txt
>
> Andrzej J. Pindor,
> University of Toronto, Computers and Communications
> pindor@utirc.utoronto.ca
>
> A solution to the symbol grounding problem proposed by Harnad requires
> giving a system both linguistic and sensorimotor capacity
> indistinguishable from those of a human. The symbols are then grounded
> by the fact that analog sensorimotor projections on transducer
> surfaces, coming from the real world objects, and successively formed
> sensory invariants of nonarbitrary shape constrain the symbol
> combinations over and above what is imposed by syntax, and tie the
> symbols to those real objects. It is argued here that the full
> sensorimotor capacity may indeed be a crucial factor, since it is
> capable of providing the symbols (corresponding to language terms) with
> a deep underlying structure, which creates a network of intricate
> correlations among them at a level of primitive symbols based in inputs
> from the transducers. On the other hand, the nonarbitrary shapes of
> sensory invariants as well as the analog nature of sensorimotor
> projections seem to be of no consequence. Grounding is then seen as
> coming from this low-level correlation structure and, once known, could
> in principle be programmed into a system without a need of
> transducers.

My reply is simple: Yes, robotic grounding consists of correlations -- correlations among (1) things in the robot’s environment, (2) the "shadows" those things cast on the robot’s transducers, (3) the robot’s internal states, and (4) what the robot can and does do in the world. But HAVING robotic capacities means INSTANTIATING those correlations, not merely ENCODING them. And that’s what grounding calls for. I could encode it all in a book, but that would be just as ungrounded as a more dynamic encoding in a computer, because it would lack the robotic capacity itself, the capacity to interact with the objects, events and states-of-affairs in the world that all the symbols -- and correlations -- are otherwise merely INTERPRETABLE as being about.

Symbol grounding requires complete autonomy from external interpretation, an unmediated CAUSAL CONNECTION between symbols and their objects (not just interpretable correlations between symbols and symbols, and between symbols and interpretations). In my model, that autonomy comes from robotic interaction capacity (T3-scale) with the objects the symbols are about; as a matter of LOGIC, that cannot be done by anything (symbols, codes, correlations) unless it is causally connected to those objects. And causal connection requires transduction. (And I don't care at all whether, at root, transduction is continuous or discrete; the point is merely that it's not just symbolic, and hence can't be done by just a digital computer.)

I will also point out below that the idea that the role of grounding is merely to "tune" internal symbolic states is homuncular and incorrect. Grounding is not a channel TO a homunculus; it is PART of the homunculus (i.e., part of ME).

> This lack of meaning of the symbols is, Harnad claims, evident for
> instance from the fact that one cannot learn Chinese from a
> Chinese-Chinese dictionary (Harnad 1993a).

My dictionary-go-round example is meant to give a flavor of WHY and HOW a symbol system is ungrounded, but it is not a demonstration. There is no way to show that it is evident that a system lacks meaning. (Not even Searle's Argument is a demonstration; Searle just shows that if you wanted to keep believing that every implementation of a T2-[Turing Test]-passing computer program understands, it would be at the price of believing either that (a) Searle would really come to understand Chinese [consciously, of course] if he just manipulated enough meaningless symbols, or, worse, that (b) manipulating meaningless symbols would generate a second, Chinese-understanding mind in Searle, of which he was not consciously aware. The other-minds problem rules out any stronger evidence than this; to know for sure whether or not it understands [indeed, whether or not anyone is home in there], you'd have to BE the system -- as Searle is, in (a).)

Nor is grounding a guarantor of meaning (CONSCIOUS meaning, that is, which is the only kind there is, by my lights, but that's another story [Harnad1993c]). A system (even T3-scale) could be grounded yet there might STILL be no meaning in there. The only thing grounding immunizes against is the objection that the connection between symbols and what they are interpretable as meaning is mediated by an external interpreter. Groundedness guarantees autonomy from external interpretation.

> To what extent are the shapes of sensorimotor projections
> 'nonarbitrary'? I will consider below several examples indicating that
> the shapes of the sensorimotor projections seem to be to a large extent
> dependent on the physical nature of the transducers, which are, in a
> sense, results of evolutionary 'accidents' (naturally optimized within
> an accessible range of physical parameters) and are then to a large
> degree arbitrary.

This is no objection. The pertinent property of the transducers is what J.J. Gibson called "affordances": There are, for example, many ways to transduce light, and many ways to get spatial cues from it to guide spatial locomotion. But the information really has to be there, in the sensory signal, and the transducer really has to pick it up. There are countless things that "afford" sittability-upon. The only invariant they share is that they are all hospitable to the shapes of our posteriors. And our posteriors might have been shaped otherwise. So what? None of this makes "being able to be sat upon" an arbitrary property. And it is properties like THAT that the robot must be able to pick up and use, in generating its T3-scale capacity.

How long that nonarbitrary shape is PRESERVED in analog form in the internal processing it undergoes (as opposed to being discretized and going into symbols) is not at issue; some tasks (like mental rotation) look as if they might be performed by preserving the shape of the sensory input right through to the motor output; others, like estimating coordinates, look as if they go quickly into symbols. The essential point is that ALL these capacities are grounded in the nonarbitrary shapes of sensorimotor icons and invariants at the transducer surface.

Another mistake in the examples below is focussing on the PHENOMENAL quality of the sensory signal, instead of the robotic capacity itself, and its (nonarbitrary) relation to the world. The nonarbitrariness is not so much concerned with what a thing LOOKS like, subjectively, but with what you can DO with it (as in the case of the invariants that subserve sittability-upon), objectively, in relation to its physical "shape" and the "shape" of the "shadow" it casts on your sensory surfaces.

> With a somewhat different chemistry the ranges of sensitivity of the
> colour cells might have been different, resulting in a different colour
> perception. One can also conceivably imagine that, had the evolution of
> the human eye gone somewhat differently, we might have ended up with a
> colour vision mechanism distinguishing three or five colours.
> Consequently, sensory projections of real objects, coming from the
> colour vision system, would have different "colour shapes", which are
> to a large extent determined by the physical nature of the
> _transducers_ and not the objects themselves.

All true. But all this means is that (a) the invariants that "afford" our actual color discrimination and identification capacity are not unique (that there is more than one way to skin the same robotic cat) or that (b) if things had evolved differently, we would have had different robotic capacities. Neither of these contradicts the nonarbitrariness of the invariants in the sensorimotor projection that the system would have to pick up in any case.

> The fact that we see the real world objects "right way up" is a result
> of the brain learning to _correlate_ shapes of sensory projections from
> the visual system with other sensory projections. One could also
> speculate that if in the distant past evolution chose a slightly
> different route we might have ended up with eyes more like those of
> insects - sensory projections of our visual system, coming from real
> world objects, would be very different and there is no reason to doubt
> that our brain would learn to deal with such a situation. We see again
> that the shapes of the sensory projections are in some sense arbitrary,
> determined by the physical nature of the transducers.

Here the focus on phenomenology has even injected some of the confusions that arise from homuncular thinking (about a little-man-in-the-head who "sees" our inputs). Of course spatial locomotion involves correlations. The correlations may not be unique ones, but they are hardly arbitrary. And their nonarbitrariness has nothing to do with what would be right-side-up to a homunculus! And, in the end, which are the correlations that really matter: It is the correlations between sensorimotor input and output, the shape of neither of which is arbitrary.

> The above examples seem to put in doubt Harnad's claim that
> "nonarbitrary shapes" of sensorimotor projections from real objects
> onto transducer surfaces are a crucial element of symbol grounding.
> The shapes of the sensorimotor projections are shown to be arbitrary to
> a large extent and it is the _correlations_ among these projections
> which appear to play a dominant role.

Correlations, yes, but between arbitrary shapes? Not at all. Shape invariants and their intercorrelations are underdetermined, no doubt (thank goodness, for if there were only one way, nature might never have had the time to find it), but hardly arbitrary (what's arbitrary is the "shape" of formal symbol tokens). And again, the pertinent thing about the correlations is their PHYSICAL REALIZATION in actual robotic interactions. The correlations, if you wish, are wider than the robot's head; they are not just correlations (between symbols) WITHIN the robot's head; to suppose they are is homuncular (and hermeneutic -- interpreting the symbols and their correlations as a homunculus's thoughts about objects).

> Harnad illustrates [the] categorization process leading to the grounding of
> the category "names" by an imaginary example of learning to distinguish
> between edible and poisonous mushrooms (Harnad 1993). It is interesting
> to note that in his example the grounding of the mushroom names
> ("mushrooms" for the edible ones and "toadstools" for the poisonous
> ones) takes place on the basis of _correlations_ between various
> sensory projections. _Shapes_ of the projection invariants do not to
> enter in any way.

But where do you think the sensory projections come from? They are shadows of the mushrooms. And the correlation is between what the robot can do with THIS kind of mushroom and not THAT kind of mushroom: Mushrooms "afford" edibility, toadstools do not. The homuncular picture (of a little man in the head doing correlations) is simply, well, wrong-headed. The edible mushroom has a nonarbitrary shape; it's shadow on the robot's transducer surfaces does too; and whatever is going on inside, the punchline has to be that the robot eats the right shapes and does not eat the wrong shapes. All that sounds pretty nonarbitrary to me, and pretty shape-dependent too. The grounding problem then becomes: how to "connect" inner symbols to the right nonarbitrary shapes, and how those connections constrain what would otherwise be just the syntactic combinatory activity of ungrounded symbols (my candidate happens to be neural nets that connect sensory projections to the symbolic names for perceptual categories by filtering out their invariants).

> [N]owhere in his arguments does Harnad convincingly show that [the]
> analog feature of the input (in the form of sensorimotor projections)
> to neural nets which do the invariant extraction is, in fact,
> essential. Any analog signal can be approximated with arbitrary
> accuracy by a digital signal. Since neural nets can have only finite
> sensitivity, whether they are fed an analog signal or a correspondingly
> finely graded digitized signal cannot matter for further processing.
> Once we accept this, these digitized signals from the transducers
> (sensorimotor projections) can be viewed as primitive symbols, in the

> same spirit as 0’s and 1’s of a Turing machine. All further processing
> can be considered as symbol manipulations which one way or another lead
> to construction of high-level symbols representing language terms
> (category names). This may very well happen with the use of neural nets
> to extract invariants from sensory projections and perhaps perform
> categorization. Since any neural net may be emulated using a suitably
> programmed digital computer, all these steps can be achieved without a
> need for analog devices.

I have repeatedly written that I have no vested interests either in whether nature is basically discrete or continuous, or in whether robot-internal processes are discrete or continuous. What I mean by "analog" is, in the end, just "physical," described by differential equations rather than implementation-independent symbol manipulation. And no matter how you discretize it, the transduction of photons is a physical process, not a computational one, and a real optical transducer will always be one kind of physical object, and a virtual transducer (a computer-simulated transducer) will be another. (At this juncture, those who have not understood this point about the physical nature of transduction usually confuse things still further by invoking virtual worlds: but those are just inputs to HUMAN SENSES -- which are, as far as I know, REAL (not VIRTUAL) transducers...)

To summarize: the critical transaction transpiring at the transducer surface is non-negotiable: Everything else from there on in may conceivably be replaceable by a digital computer (in principle, if not in practice), but sensorimotor transduction itself is not; yet that is the critical locus and causal vehicle of the "correlations" that are really at issue here.

About higher-order versus lower-order symbols, see Harnad (1993c). The short answer is that higher-order symbols have about as good a chance of breaking out of the hermeneutic circle as an acrostic does: If the lower-order symbols are just ungrounded squiggles and squoggles, then that’s all the higher-order ones are, and vice-versa.

> The above analysis suggests that full robotic capacity of a system
> might provide high-level symbols with a deeper structure based in
> correlations among the primitive symbols, the sources of which are
> inputs from sensorimotor transducers. Symbol grounding would then be
> achieved by the presence of such an underlying structure, which would
> give the symbols a much richer (and more intricate) set of
> relationships than can be offered by a (single-language) dictionary.
> These relationships mirror the experiences of interacting with the real
> world, making the symbols effective in such interactions and
> justifying the claim that the symbols are grounded.

This does not show how to break out of the dictionary-go-round. Be it ever so "rich" and intercorrelated, it still sounds like the cacophony of meaningless symbols to me (unless I project an interpretation onto it).

Here comes the Cheshire Cat’s Grin (or, to mix metaphors, the ungrounded symbols, hanging from a skyhook): Sure there are lots of correlations within robot-internal states that go into delivering T3 power. But in and of themselves, they deliver nothing. (It would be an instance of the mirror-image of the virtual-world confusion alluded to above to equate the inner states of a REAL

Helen-Keller-Stephen-Hawking -- which still consist mostly of transduction, albeit damaged transduction -- with the inner states of a digital computer, which could never do ON ITS OWN what even H-K-S-H can still do; the computer could only do so if you ADDED sensorimotor transducers to it (transducers that, I repeat, H-K-S-H does not lack).

Here is an intuition pump that has sometimes worked (I've even posted it here a few times): "I AM a (sensorimotor) transducer." I may not be JUST a sensorimotor transducer (I may even be part digital computer), but something that is JUST digital computer is not, and cannot be, me (or anybody), no matter what correlations it encodes, because the RIGHT correlations for implementing ME include the (causal capacity for) sensorimotor interaction as an ESSENTIAL component. To suppose otherwise is just to be seduced by homuncular hermeneutics on symbols (their intercorrelations among themselves and their systematic "correlation" with what they are interpretable as being about).

And no, sensorimotor grounding is not merely the standard way that the inner correlations happen to get SET (by evolution and learning) in real time. It is the implementation of the sensorimotor capacity itself, in the here-and-now, that makes me me (a T3 robot could have sprung fully formed out of the head of Zeus, and it would STILL be grounded if it did indeed have T3 capacity; its "inner correlations," severed from its sensorimotor apparatus, on the other hand, would be nothing, no one, no better than a Chinese/Chinese dictionary [which has plenty of correlations too, by the way]).

> It is nevertheless worth pointing out that there does not seem to be a
> reason why the underlying structure discussed above, once established,
> could not be built (programmed) into a symbolic system, without a need
> to give the system the full robotic capacity. Such a system would be
> capable of passing the TT and should perhaps also be considered to
> possess understanding of the language it uses.

You could build even a T3 robot without any real-time history of any kind, and its internal symbols would be grounded (interpreter-independent) in virtue of its T3 (CAUSAL) capacity. But if you just had a T2 symbol system that was interpretable as performing the "right" correlations -- and even if that symbol system had previously been a proper part of the T3 robot itself -- the symbols in that system would nevertheless be just as ungrounded as the intergalactic vacuum (or the robot's assembly-manual).

To put it another way: A T3 system, sitting here, idling, is still perfectly grounded, in virtue of its REAL sensorimotor capacity, a kind of physical potential it really has. On the other hand, a pure symbol-crunching module, chugging away, with its symbols ever so systematically interpretable, and its correlations all raring to go, if only it were reinserted in a T3 robot, is ungrounded, because it has NO sensorimotor capacity. "Capacity" may mean "potential," but the "potential capacity" of such a system is getting TOO potential; in reality, it is T3 impotent (otherwise, by that token, a single flip-flop would already be a potential computer, hence mind, too!)

> There is one more aspect of the grounding problem, as discussed above,
> which requires mentioning. There are situations when we deal with
> concepts defined solely using language, without a reference to
> sensorimotor projections from real world objects. Such situations
> arise, for instance, in the case of mathematics. If we consider

> abstract set theory or abstract group theory, we define objects (sets,
> group elements) purely syntactically and then proceed to draw all
> possible conclusions concerning the consequences of these definitions.
> In spite of the fact that the symbols we manipulate do not require
> grounding in sensorimotor projections from real world objects, and the
> manipulations depend only on shapes of these symbols (which are
> completely arbitrary), we do talk about "understanding" mathematics
> (abstract set theory, abstract group theory, etc.). It is clear that
> understanding in this case means a knowledge of (or ability to deduce)
> _correlations_ among symbols of increasing complexity, arising from
> definitions of basic symbols from which these higher level symbols are
> constructed.

This is too big a topic to take up here. The short answer is that I believe that all higher-order, abstract terms are grounded in lower-order ones (that are grounded in... etc,) and ultimately in sensorimotor categories. (See my discussion of the "peekaboo unicorn" in Harnad 1992a.) Grounding is recursive, in mind-building as much as in skyscraper building. Nothing hangs from a skyhook; you have to get there bottom-up.

That said, it IS possible to "do" mathematics as pure symbol manipulation (along the lines that the formalists would say ALL of mathematics goes). Everyone who uses a "cookbook formula" mechanically, without understanding what he’s doing, is doing that. But then you are in Searle’s Chinese Room, and there really IS no understanding or meaning going on. (I doubt that that’s what the formalists had in mind; certainly Roger Penrose does not think that’s all that’s going on in the mathematician’s mind.)

> In conclusion, it is argued above that even though two aspects of
> Harnad’s model for symbol grounding seem unjustified:
>
> - shapes of sensorimotor projections from real objects onto transducer
> surfaces do not appear to be relevant and hence cannot play a role in
> restricting symbol combinations;

This is answered by the affordance and sensorimotor correlation discussion above. And grounding could not be just a matter of restricting symbol combinations, otherwise one could have said a priori that all grounding required was the right set of syntactic rules, hence there was no symbol grounding problem. The symbol grounding problem is the problem of getting the external interpreter out of the causal loop that connects symbols to what they are about.

> - importance of the analog nature of the sensorimotor projections, fed
> subsequently to neural nets for invariant feature extraction, is not
> apparent (there are reasons to think that these projections might just
> as well be digitized, leaving us with pure symbol manipulations);

You can digitize what comes AFTER the transduction, internally, but you can’t digitize the sensorimotor transduction itself.

> the main idea of the model - TTT capacity - may be crucial
> for symbol grounding. It may be the combinations of various sensorimotor
> experiences with real objects which lead to the formation of a deep
> structure underlying the high-level symbols which provides
> (epistemological) meaning of language terms.

No, grounding is not just the education of an inner symbolic homunculus. Nor is it just providing the right data to tune a pure symbol cruncher. We ARE sensorimotor transducers (at least in part), and our symbols are grounded in our ACTUAL sensorimotor capacities. So a grounded symbol system is fated to be more than just a symbol system. I'm betting the hybridism is analog/neural-net/symbolic, with the lion's share of the load borne by the analog component.

> There also appears a possibility that if a symbolic system works
> on the basis of digitized inputs, corresponding to sensorimotor
> projections coming from transducers, as basic symbols, it might
> possess understanding without TTT capability. Possibility of
> ascribing understanding to a purely symbolic system seems in
> accordance with using the term "understanding" in the case of
> abstract mathematics, where (mathematical) terms used are
> described verbally only, without recourse to the full sensorimotor
> capacities of a human being.

But, like Helen Keller and Stephen Hawking, abstract mathematicians are themselves grounded sensorimotor systems, so nothing follows about ungrounded symbol systems from any of this. And we are not interested in what properties can be ASCRIBED to systems (by interpreting them) but in what properties they really HAVE, autonomously, over and above their systematic interpretability. EVERYTHING, besides (a) having the properties it really has, is also (b) systematically interpretable as having those properties; but virtual systems (ungrounded symbol systems) have only (b) -- AS virtual systems: we all know what a digital computer REALLY is, independent of what it might be interpretable as simulating.

Beware the Cheshire Cat's Grin.

Stevan Harnad

The following files are retrievable from directory pub/harnad/Harnad on host princeton.edu

Harnad, S. (1987) The induction and representation of categories. In: Harnad, S. (ed.) (1987) Categorical Perception: The Groundwork of Cognition. New York: Cambridge University Press.

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25. FILENAME: harnad89.searle

Harnad, S. (1990) The Symbol Grounding Problem. Physica D 42: 335-346. FILENAME: harnad90.sgproblem

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54. FILENAME: harnad91.otherminds

Harnad, S. (1992a) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag. FILENAME: harnad92.symbol.object

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on the Virtual Mind. Minds and Machines 2(3) 217-238. FILENAME: harnad92.virtualmind

Harnad, S. (1992b) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. SIGART Bulletin 3(4) (October) pp. 9 - 10 FILENAME: harnad92.turing

Harnad, S. (1993a) Artificial Life: Synthetic Versus Virtual. Artificial Life III (Santa Fe, June 1992) (to appear) FILENAME: harnad93.artlife

Harnad, S. (1993b) The Origin of Words: A Psychophysical Hypothesis. Presented at Zif Conference on Biological and Cultural Aspects of Language Development. January 20 - 22, 1992 University of Bielefeld; to appear in Durham, W & Velichkovsky B (Eds.) "Naturally Human: Origins and Destiny of Language." Muenster: Nodus Pub. FILENAME: harnad93.word.origin

Harnad, S. (1993c) Grounding Symbols in the Analog World with Neural Nets. Think 2: 12 - 78 (Special Issue on "Connectionism versus Symbolism" D.M.W. Powers & P.A. Flach, eds.). FILENAME: harnad93.symb.anal.net

Harnad, S. (1993d) Symbol Grounding is an Empirical Problem: Neural Nets are Just a Candidate Component. Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society. NJ: Erlbaum FILENAME: harnad93.cogsci

Harnad, S. (1993e) Problems, Problems: The Frame Problem as a Symptom of the Symbol Grounding Problem. PSYCOLOQUY 4(34) frame-problem.11. FILENAME: harnad93.frameproblem

Harnad, S. (1993f) Grounding Symbolic Capacity in Robotic Capacity. In: Steels, L. and R. Brooks (eds.) The "artificial life" route to "artificial intelligence." Building Situated Embodied Agents. New Haven: Lawrence Erlbaum FILENAME: harnad93.robot

Harnad, S, (1994) Does the Mind Piggy-Back on Robotic and Symbolic Capacity? To appear in: H. Morowitz (ed.) "The Mind, the Brain, and Complex Adaptive Systems. FILENAME: harnad94.mind.robot

Harnad, S. (1994) Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. Artificial Life 1(3): (in press) FILENAME: harnad94.artlife2

E’PUR’ NON SI MUOVE

> From: rickert@mp.cs.niu.edu (Neil Rickert)
> Subject: Re: Symbol grounding
> Organization: Northern Illinois University
> Date: Fri, 28 Jan 1994 18:34:53 GMT

>
> We may be able to talk of a Turing machine as a formal system we dream
> up in our minds, but a computer is a practical device built from
> transducers.

Of course I'm aware of this. Any physical system is a piece of physics, and to the extent that energy exchange is involved, it is a transducer. A digital computer is a bunch of transducers. The relevant point about digital computers (construed as computers rather than just mysteriously flip-flopping devices) is that they perform COMPUTATION, and computation is the manipulation of objects with arbitrary shapes, based on rules (syntactic ones) operating only on the basis of those arbitrary shapes in such a way that the symbols and manipulations are SYSTEMATICALLY INTERPRETABLE in some way. e.g., as numerical calculations, or language (syntactic gibberish might be a special case of uninterpretable computation, but it would also be a trivial case).

And -- now here comes the critical part -- the physical implementation of the computation is IRRELEVANT. The computation is IMPLEMENTATION-INDEPENDENT. It could have been performed in countless, radically different ways (all of them physical, each involving transduction, but the specifics of the transduction are in every case irrelevant).

It is THIS property, implementation-independence, that Searle's Argument capitalizes on: For if "computationalism" were true, i.e., if it were true that "Cognition is Just a Form of Cognition," then, being computation, cognition would have to be implementation-independent (and so Searle -- or somebody in there -- would have to understand Chinese, because he was just an another implementation of the same, T2-passing computation).

So the answer is that the transducers that implement a computation are irrelevant. The transduction *I* was referring to was the sensorimotor transduction done by the portion of a T3-scale robot implementing its T3-interactions with the world. Those transductions are NOT irrelevant -- either to T3, obviously, or to having a mind, by my lights; nor are they computations. And a digital computer (alone) cannot perform them: it's the wrong type of physical device for that.

And, as I said before, IF, mirabile dictu, a digital computer hooked up to a simple set of peripherals COULD pass T3, the only thing that would follow would be that computer + peripherals had a mind, but computer alone still would not.

I don't believe for a moment that sensorimotor transduction would be that trivial; nor that most of the heavy work of implementing cognition is done by a computational module hooked up to peripherals (the real brain is certainly doing mostly sensorimotor transduction). But even if that were true, it would STILL be false that cognition is just a form of computation, and digital computers still would not have minds. It would also still follow that symbol grounding does not amount, as Pindor's original commentary suggested, to just a way of getting the right data into a digital computer.

> There are no symbols [in a computer]. The symbols are in our human
> readable computer programs before they are translated by the compiler.
> The computer itself deals only with electrical currents. The symbols
> are part of a human interpretation of the computer, and are not
> themselves an inherent part of computer operation. Conventional
> computer programs seem to be highly symbolic only because that suits

> the uses to which we humans wish to put computers.

A digital computer is the dynamic implementation of a symbol system; the program on paper is just a static implementation (to the extent that it is an implementation at all). But it is true of both that certain physical objects -- on paper, they are the (binary level) symbol tokens, in the computer, they are the states of the flip-flops -- are implementing certain formal objects, and that the physical manipulations of those physical implementations of those formal objects according to the shape-governed (syntactic) rules that are being followed can be given a systematic interpretation (as arithmetic, or logic, or language). That's not trivial. It requires an uncanny correspondence between physical states and certain systematic interpretations that they can be given. That is the power of formal computation (and, when implemented physically, the power of digital computers).

It is also true, however, that just as the "shape" of the notational system in the paper program is irrelevant to the computation, so the details of the physical implementation of the digital computation are irrelevant to the computation. So, to the extent that the hypothesis in question has to do with whether or not X (whatever it is) is indeed JUST A FORM OF COMPUTATION, I *can* restrict my attention to the symbol system a digital computer is implementing, ignoring any other facts about its physical make-up.

Besides, whatever computers have had going for them as potential minds, it didn't have anything to do with the transducers they were made of, but with the computations they could perform; those computations are ungrounded, hanging from an interpretational skyhook; our thoughts (whatever they are) are not.

This is all a very tricky area to think about, something I've often dubbed "The Hermeneutic Hall of Mirrors," where you can easily get lost in the reflections of the interpretations that you have yourself projected onto a symbol system, forgetting that you are their sole source (Harnad 1990a,b).

> I do not mean by the above to imply that a keyboard, display terminal,
> and disk drive, are adequate peripheral equipment to solve the symbol
> grounding problem. Indeed, I agree with you that this would not
> suffice. What I am suggesting is that there is something wrong with
> your explanation as to why a simple desktop computer would not
> suffice.

The short answer is that a digital computer plus a keyboard display terminal cannot pass T3. THAT's why it's inadequate. (In principle, though, they would be enough to allow an ingenious modeller to successfully second-guess everything it WOULD take to build a winning T3 robot: Yet even THAT [running] program would not have a mind, any more than a successful planetary motion simulation [in a virtual solar system] would have motion: Only the actual robot (and solar-system) would [Harnad 1993].)

> The digital computer is digital only in the sense that explanations in
> terms of digital computation provide the best descriptions of its
> operation. But in terms of what is happening it is a physical analog
> device through and through. The questions about the brain are not
> whether neurons are intrinsically digital devices, but are about
> whether digital computational explanations are useful in describing the
> activity of the neural system. We do not currently know enough to

> definitively answer these questions.

I think we already know enough to say that digital computational explanations ARE useful in describing (some of) the activity of the neural system. That's what Searle called "Weak AI" (and what I would call "Weak Computationalism"). But what we can also say with some certainty is that "Strong Computationalism" is false: Cognition is NOT just a Form of Computation. Hence no digital computer (alone) could have a mind.

> I am going to insist that a digital computer IS a sensorimotor
> transducer. It might be reasonable to ask if it is the right kind of
> transducer. But to argue that it is not a transducer is to confuse the
> real practical physical computer sitting on your desk with a formal
> abstract Turing machine existing only in the mind of the
> mathematician.

A digital computer is not the right kind of transducer because it cannot pass T3; for that you need the sensorimotor transducers of a T3 robot. (For similar reasons, I would say that a digital computer is not a SENSORIMOTOR transducer at all, but let's not quibble.)

> Let me stipulate that I agree with Harnad as to the importance of symbol
> grounding. For example I have often suggested in this newsgroup that
> the Robot Reply is the correct response to Searle's CR argument. I take
> the TTT as Harnad's version of the Robot Reply.

I hate to quibble about this too, but the various versions of the "Robot Reply" that I have read are all unprincipled ones, easily refuted by Searle, because they have neither (1) realized nor (2) motivated, nor (3) bitten the bullet and assumed the consequences of committing themselves to: T3 and the essential role of sensorimotor transduction in cognition (i.e., exactly what's under discussion here). It was for that reason that I wrote in Harnad (1989):

Searle (1980a) has dubbed the various prior replies to his argument the "Systems Reply," the "Robot Reply," etc. One is tempted to call this one the "Total Reply," for reasons that will become apparent. Lest this be misunderstood as imperialism, however, perhaps it should rather just be referred to as the "Robotic Functionalist Reply."

The belief that symbol grounding via T3 interactions is just (1) Nature's normal way -- or (2) the theorist's best or only way -- to tune a symbol system with the right data is, in my view, incorrect; and, far from being a solution to the symbol grounding problem, it is a symptom of it.

Stevan Harnad

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Harnad, S. (1990a) Against Computational Hermeneutics. (Invited commentary on Eric Dietrich's Computationalism) Social Epistemology 4: 167-172.

Harnad, S. (1990b) Lost in the hermeneutic hall of mirrors. Invited Commentary on: Michael Dyer: Minds, Machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence 2: 321 - 327.

Harnad S. (1993) Discussion (passim) In: Bock, G.R. & Marsh, J. (Eds.) Experimental and Theoretical Studies of Consciousness. CIBA Foundation Symposium 174. Chichester: Wiley

----

> Date: Fri, 28 Jan 1994 14:55:12 -0800
> From: ken@holonet.net (Ken Easlon)
>
> We could conceivably build a SHRDLU style robot in a simulated
> environment, down to the details of simulating transduction. In
> this kind of set up, the "physical" world of the robot would be
> implemented by modules that simulate the laws of physics, and the
> robot's internal states might be implemented by modules that
> emulate a human-like mind.
>
> Harnad's distinction between the robot's internal states and the
> physical-like qualities of the (simulated) transducers and
> environment would still apply in this simulation paradigm.
>
> However, in my view, his refusal to "negotiate" the possibility of
> using computers in this realm is philosophically short sighted.

What's non-negotiable is not whether one "uses" computers, but what it takes to (a) pass T3 and (b) have a mind. I have no say over (b), but a virtual robot in a virtual world does not pass T3. See the references below.

Stevan Harnad

The following files are retrievable from directory pub/harnad/Harnad on host princeton.edu (citation is followed by FILENAME and, where available, ABSTRACT):

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on the Virtual Mind. Minds and Machines 2(3) 217-238. FILENAME: harnad92.virtualmind ABSTRACT: When certain formal symbol systems (e.g., computer programs) are implemented as dynamic physical symbol systems (e.g., when they are run on a computer) their activity can be interpreted at higher levels (e.g., binary code can be interpreted as LISP, LISP code can be interpreted as English, and English can be interpreted as a meaningful conversation). These higher levels of interpretability are called "virtual" systems. If such a virtual system is interpretable as if it had a mind, is such a "virtual mind" real? This is the question addressed in this "virtual" symposium, originally conducted electronically among four cognitive scientists: Donald Perlis, a computer scientist, argues that according to the computationalist thesis, virtual minds are real and hence Searle's Chinese Room Argument fails, because if Searle memorized and executed a program that could pass the Turing Test in Chinese he would have a second, virtual, Chinese-understanding mind of which he was unaware (as in multiple personality). Stevan Harnad, a psychologist, argues that Searle's Argument is valid, virtual minds are just hermeneutic overinterpretations, and symbols must be grounded in the real world of objects, not just the virtual world of interpretations. Computer scientist Patrick Hayes argues that Searle's Argument fails, but because Searle does not really implement the program: A real implementation must not be homuncular but mindless and mechanical, like a computer. Only then can it give rise to a mind at the virtual level. Philosopher Ned Block suggests that there is no

reason a mindful implementation would not be a real one.

Harnad, S. (1993) Artificial Life: Synthetic Versus Virtual. Artificial Life III (Santa Fe, June 1992) (to appear) FILENAME: harnad93.artlife ABSTRACT: Artificial life can take two forms: synthetic and virtual. In principle, the materials and properties of synthetic living systems could differ radically from those of natural living systems yet still resemble them enough to really be alive if they are grounded in the relevant causal interactions with the real world. Virtual (purely computational) "living" systems, in contrast, are really just ungrounded symbol systems that are systematically interpretable as if they were alive; in reality they are no more alive than a virtual furnace is hot. Virtual systems are better viewed as "symbolic oracles" that can be used (interpreted) to predict and explain real systems, but not to instantiate them. The vitalistic overinterpretation of virtual life is related to the animistic overinterpretation of virtual minds and is probably based on an unconscious (and possibly erroneous) intuition that living things have actual or potential mental lives.

Harnad, S. (1993) Grounding Symbols in the Analog World with Neural Nets. Think 2: 12 - 78 (Special Issue on "Connectionism versus Symbolism" D.M.W. Powers & P.A. Flach, eds.). FILENAME: harnad93.symb.anal.net ABSTRACT: The predominant approach to cognitive modeling is still what has come to be called "computationalism," the hypothesis that cognition is computation. The more recent rival approach is "connectionism," the hypothesis that cognition is a dynamic pattern of connections and activations in a "neural net." Are computationalism and connectionism really deeply different from one another, and if so, should they compete for cognitive hegemony, or should they collaborate? These questions will be addressed here, in the context of an obstacle that is faced by computationalism (as well as by connectionism if it is either computational or seeks cognitive hegemony on its own): The symbol grounding problem. WITH ACCOMPANYING COMMENTARIES AND RESPONSES: harnad93.symb.anal.net.boyle harnad93.symb.anal.net.bringsjord harnad93.symb.anal.net.dietrich harnad93.symb.anal.net.dyer harnad93.symb.anal.net.fetzer harnad93.symb.anal.net.hayes harnad93.symb.anal.net.honavar harnad93.symb.anal.net.kentridge harnad93.symb.anal.net.maclennan harnad93.symb.anal.net.mcdermott harnad93.symb.anal.net.powers harnad93.symb.anal.net.roitblat harnad93.symb.anal.net.searle

Harnad, S. (1993) Grounding Symbolic Capacity in Robotic Capacity. In: Steels, L. and R. Brooks (eds.) The "artificial life" route to "artificial intelligence." Building Situated Embodied Agents. New Haven: Lawrence Erlbaum FILENAME: harnad93.robot

Harnad, S, (1994) Does the Mind Piggy-Back on Robotic and Symbolic Capacity? To appear in: H. Morowitz (ed.) "The Mind, the Brain, and Complex Adaptive Systems. FILENAME: harnad94.mind.robot ABSTRACT: Cognitive science is a form of "reverse engineering" (as Dennett has dubbed it). We are trying to explain the mind by building (or explaining the functional principles of) systems that have minds. A "Turing" hierarchy of empirical constraints can be applied to this task, from t1, toy models that capture only an arbitrary fragment of our performance capacity, to T2, the standard "pen-pal" Turing Test (total symbolic capacity), to T3, the Total Turing Test (total symbolic plus robotic capacity), to T4 (T3 plus internal [neuromolecular] indistinguishability). All scientific theories are underdetermined by data. What is the right level of empirical constraint for cognitive theory? I will argue that T2 is underconstrained (because of the Symbol Grounding Problem and Searle’s Chinese Room Argument) and that T4 is overconstrained (because we don’t know which neural properties, if any, are relevant). T3 is the level at which we solve the "other minds" problem in everyday life, the one at which evolution operates (the Blind Watchmaker is no mind-reader either) and the one at which symbol systems can be grounded in the robotic capacity

to name and manipulate the objects their symbols are about. I will illustrate this with a toy model for an important component of T3 -- categorization -- using neural nets that learn category invariance by "warping" similarity space the way it is warped in human categorical perception: within-category similarities are amplified and between-category similarities are attenuated. This analog "shape" constraint is the grounding inherited by the arbitrarily shaped symbol that names the category and by all the symbol combinations it enters into. No matter how tightly one constrains any such model, however, it will always be more underdetermined than normal scientific and engineering theory. This will remain the ineliminable legacy of the mind/body problem.

Harnad, S. (1994) Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. Artificial Life 1(3): (in press) FILENAME: harnad94.artlife2 ABSTRACT: Both Artificial Life and Artificial Mind are branches of what Dennett has called "reverse engineering": Ordinary engineering attempts to build systems to meet certain functional specifications, reverse bioengineering attempts to understand how systems that have already been built by the Blind Watchmaker work. Computational modelling (virtual life) can capture the formal principles of life, perhaps predict and explain it completely, but it can no more BE alive than a virtual forest fire can be hot. In itself, a computational model is just an ungrounded symbol system; no matter how closely it matches the properties of what is being modelled, it matches them only formally, through the mediation of an interpretation. Synthetic life is not open to this objection, but it is still an open question how close a functional equivalence is needed in order to capture life. Close enough to fool the Blind Watchmaker is probably close enough, but would that require molecular indistinguishability, and if so, do we really need to go that far?

Date: Sat, 29 Jan 94 15:26:34 EST From: "Stevan Harnad" To: sharnad@life.jsc.nasa.gov Subject: Bringsjord on Proof, Possibility and Mind

To: Symbol Grounding Discussion Group

> From: Selmer Bringsjord (brings@rpi.edu)
> Date: Fri, 28 Jan 94 17:52:56 EST
> To: harnad@Princeton.EDU
>
> Hello Stevan, Enjoyed the ai.comp.phil exchange. Perhaps I could post
> two things: my soon-to-be-posted :-) (on sym gr list) argument that
> your conditional is false (which I've lost again; do you keep stealing
> that from my space here :-)?), and an argument that it's physically
> possible that I be a brain in a vat w/o sensors and effectors. Such a
> physical possibility undercuts your position, as I understand it. There
> is also a third possibility: Brentano/ Chisholm/Bringsjord arguments
> that certain intentional states can't be explained in causal
> agent-environment transactional terms. Yours, selmer

Dear Selmer, I've re-read your September message ("Harnad on the Brink") and am posting it to the SG Group below (apologies for the long delay), noting that I have no reply, as my reply would only be a repetition of my previous comments, already abundantly re-quoted in your posting itself. But at this point I also have to invoke cloture on our own back-and-forth on this topic (proof, possibility, etc.). I think we have had our respective says and we should now let the rest of the cognitive community draw its own conclusions. Best wishes, Stevan

----

From: Selmer Bringsjord Date: Wed, 8 Sep 1993 22:21:52 -0400

Feature Presentation:

PART I: WHY HARNAD'S POSITION ON TURING TESTING IS FALSE

Background Music:

PART II: IF YOU'RE GONNA ARGUE ABOUT WHAT IS AND ISN'T A PROOF, SAY WHAT 'PROOF' MEANS

PART III: IN DEFENSE OF MY ARGUMENT AGAINST AI FROM FREE WILL, AND HARNAD ON THE BRINK

PART I: WHY HARNAD'S POSITION ON TURING TESTING IS FALSE

Your now-explicit position on Turing Testing:

>
>sh> The negative version I do endorse is:

"If X is T3-indistinguishable from someone with a mind, one has no nonarbitrary reason for doubting that X has a mind when told that X is a machine."

Wonderful. We've got your position now. Let's see why it's false -- or, to put it in terms you prefer, lets see why your position is very implausible.

Start by recalling that the T2 version of your conditional -- 'H,' as I'll call it -- is, by your own admission, false. That is,

T If X is T2-indistinguishable from someone with a mind, one has no nonarbitrary reason for doubting that X has a mind when told that X is a machine.

is false -- for reasons that are by now familiar: Conduct a thought- experiment in which T's antecedent is true but it's consequent is false. More specifically, suppose that Millie is T2-indistinguishable from someone with a mind. To make this a bit more vivid, suppose that roboticist Smith carries a computer around with him which is T2-indistinguishable from someone with a mind, but suppose that Smith cloaks this computer in some way (with clothes, wig, beard, a cape over a form in a wheelchair -- use your imagination) and refers to it as 'Millie.' So, you carry out conversations with Millie for years. And then one day Smith says, "Aha! Watch this!" And he pulls off the cloak, and there sits a computer with a good voice synthesizer, etc.

There are principled reasons for doubting that Millie has a mind, as you heartily agree. If I understand you correctly, we can take our pick from at least the serendipity argument and CRA. If you go with serendipity, you say to Smith: "Look, this is just a bunch of symbols swimming around inside a box. There's no person in the picture. It's easy enough to show that there's no principled connection between the symbols working nice and there being a person in the picture, for consider a situation [spelled out as follows...; see (Bringsjord, forthcoming-b)] wherein Millie's program is a lucky random sentence generator based on technology students master in the first few weeks of Intro. to AI. If you go with CRA, you could say to Smith: "Look, this is just a bunch of symbols swimming around inside a box. There's no person in the picture. It's easy enough to show that there's no principled connection between the symbols working nice and there being a person in the picture, for consider a situation [spelled out as follows...; see (Bringsjord, 1992, Chapter V: Searle)] wherein Jones implements, in Searlean style, a rock-and-box version of the program involved, and doesn't have any of the mental states while doing so which we've been ascribing to Millie."

Alright, now along comes Harnad with

H If X is T3-indistinguishable from someone with a mind, one has no nonarbitrary reason for doubting that X has a mind when told that X is a machine.

How do we show H false? Let's follow the recipe followed above w.r.t. T: Conduct a thought-experiment in which H's antecedent is true but it's consequent is false. More specifically, begin by supposing that X is T3-indistinguishable from someone with a mind. To make this a bit more vivid, suppose that roboticist Smith escorts a robot around with him which is T3-indistinguishable from someone with a mind, but suppose that Smith cloaks this robot in some way (with synthetic skin that passes for the real thing, artificial eyes that look human, etc. -- use your imagination) and refers to it as 'Willie.' So, you carry out conversations with Willie for years; *and* you play baseball (and the like) with Willie for years. And then one day Smith says, "Aha! Watch this!" And he trears off the skin over Willie's forehead, and therein sits a computer busy humming, etc.

So far so good. How do we get the falsity of your consequent built into the thought-experiment? Well, that's a piece of cake. Let's begin by reminding ourselves of what we need. We need the Willie scenario to include the falsity of

One has no nonarbitrary reason for doubting that Willie has a mind.

Let's instantiate this to me. That seems harmless enough. So we need the scenario to include the falsity of

Selmer has no nonarbitrary reason for doubting that Willie has a mind.

Here's how it works. Selmer can take his pick from at least the serendipity argument prime and CRA prime. If Selmer goes with serendipity', he says to Smith: "Look, Willie is still just a bunch of symbols swimming around inside a box. There's no person in the picture. It's easy enough to show that there's no principled connection between the symbols and sensors/effectors working nice and there being a person in the picture, for consider a situation [spelled out as follows...; see (Bringsjord, forthcoming-b)] wherein Willie's program is a *composite* program: a lucky random sentence generator *and* a lucky program for processing information received via sensors *and* a lucky program for processing information sent to effectors -- all code that's based on technology students master in the first few weeks of Intro. to AI."

If you go with CRA', Selmer can say to Smith: "Look, Willie is still just a bunch of symbols swimming around inside a box. There's no person in the picture. It's easy enough to show that there's no principled connection between the symbols and sensors/effectors working nice and there being a person in the picture, for consider a situation (spelled out as follows...) wherein Jones implements a rock- and-box version of the composite program (the one that handles linguistic performance, information from sensors, and information sent to effectors) involved, and doesn't have any of the mental states while doing so which we've been ascribing to Willie."

Note that my reasons only need to be nonarbitrary. You can tell me that these reasons don't carry the day, you can tell me that these reasons aren't compelling, you can tell me that they won't worry aspiring person-builders. But one thing you can't tell me is that my reasons are arbitrary. They're not. For they're based quite prudently on reasons you yourself concede to be not only principled but compelling.

PART II: IF YOU'RE GONNA ARGUE ABOUT WHAT IS AND ISN'T A PROOF, SAY WHAT 'PROOF' MEANS

>sb> One can't simply *assume* that it's only what aliens can
>sb> *do* that guides our judgement, for we may be able to sit down in
>sb> our armchairs and reason things out well in advance of their
>sb> arrival. If we hear in advance, for example, that the Nebulon race
>sb> is composed of beings who are physical implementations of finite
>sb> state automata, we can immediately deduce quite a bit about
>sb> the cognitive capacity of Nebulons.

Stevan replied directly with:

>
>sh> It MIGHT be possible in principle to prove that certain kinds of
>
>sh> T3-passers (Turing-indistinguishable from us in their robotic and
>
>sh> symbolic performance capacities) can't have minds, but if so, I
>
>sh> haven't a clue how such a proof might run.

The example of the Nebulons serves as a counter-example to the proposition that you apparently affirmed earlier, viz., that one can only go by what creatures *do* when trying to establish things about their cognition. I donUt know whether or not the Nebulons have minds (heck, they're sketchy fictional chaps!); I know that since the Nebulons are FSAs they canUt X, where X would be tied to automata more powerful than FSAs.

>
>sh> For, even in the
>
>sh> obvious cases (cheating by remote human telemetric control of
>
>sh> the robot, or a robot that happens to do it all purely by chance),
>
>sh> the cases one would want to reject a priori, the logical
>
>sh> POSSIBILITY that the candidate has a mind anyway cannot be
>

>sh> eliminated.

No. You’re saying that the following two propositions are inconsistent (variables arbitrary):

(1) It’s logically possible that machine M have a mind.

(2) There exists a proof that M doesn’t have a mind.

That’s the claim you’re making: (1) and (2) are inconsistent. So show me the argument for this claim -- which doesn’t rely upon your private definition of proof! Given some good candidates for what you mean by ’have a mind,’ and given the conceptual level you prefer in these matters, I can prove the *negation* of your claim:

Proof. A proof is a chain of reasoning P1, P2, P3, ..., Pn formally valid in some logical system such that a significant number of those in the community involved affirm Pn on the strength of P1-Pn-1 -- my def. from previous comm. Consider, then, this chain of reasoning:

If M can’t have qualia, M can’t have a mind. If a machine can have qualia, then qualia can be fully specified in the hybrid physics-computer science language of AI/Cog Sci. Qualia can’t be fully specified in the hybrid physics-computer science language of AI/Cog Sci. So a machine can’t have qualia. Therefore M can’t have a mind.

This reasoning is formally valid in the propositional calculus. Furthermore, a significant number of those in the community affirm M’s no-mindedness on the basis of the reasoning here involved [Searle himself, in affirming Frank Jackson’s (and Nagel’s, and KripkeUs) attack on physicalism from qualia, comes awfully close: see (Searle 1992, p. 117-8); (Bringsjord, in progress, forthcoming-a).]

It remains to be shown that it’s logically possible that M have a mind; but that’s easy: To have a mind it suffices to be associated with a mind in the Cartesian sense. It’s logically possible that Cartesian substance (or agent) dualism is true. It’s logically possible that M is associated with a Cartesian mind. Therefore, it’s logically possible that M has a mind. QED

And of course it’s even worse for you, because the scenario we’ve just gone through can be generalized, as follows. You assume that the following two propositions are inconsistent:

(1’) It’s logically possible that P.

(2’) There exists a proof that ~P.

To say that two propositions Q and R are inconsistent is to say that it’s not logically possible that [Q & R]. So to prove the negation of your claim one need only show that it’s logically possible that [(1’) & (2’)]. But it’s easy enough to describe a coherent situation which establishes this: Let (2) be verified by the fact that agents A1 to An affirm [If Q then ~P, Q, therefore ~P]. Then (2) follows. Suppose also that P can be coherently described so that from it a contradiction can’t be derived.

>
>sh> The same is true of Searle’s argument that a T2-
>
>sh> passing computer would not have a mind: No proof, it only

>
>sh> suggests that it’s extremely unlikely.

Isn’t this ironic? Two people sold on CRA who nonetheless end up arguing about it -- hmm. Look, it just never occurred to me to understand CRA probabilistically. It sounds like it never occurred to you to understand it otherwise! IUd wager that thatUs probably true of any substantive philosophical argument youUve ever seen -- proofs for GodUs existence, arguments for and against infanticide, etc. Most of the arguments that are any good in this game are deductive. Deductive arguments which are any good are demonstrably valid (i.e., it’s demonstrable that it’s impossible that the premises be true while the conclusion isn’t). *The only known way to establish validity is to formalize the argument.*

>
>sh> I don’t see why you want to turn these epistemic and
>
>sh> methodological questions into matters
>
>sh> for PROOF, rather than mere arguments for plausibility on the
>
>sh> evidence.

No, no, I say right off the bat in *What Robots Can & Can’t Be* that all I want are formally valid arguments with plausible premises -- nothing more. (And, as I’ve just said, you can’t get formal validity w/o the formal.) ***But the deeper point is that an acceptable definition of proof implies (along with logic, mathematics and technical philosophy as theyUre practiced), that thereUs not that a big difference between what IUm about and a proof***. The essence of technical philosophy is to reduce, at least in large part, a thorny clash of intuitions to formal questions. What I do is either the best or worst of technical philosophy, *depending on one’s prior prejudices* (as a commentator once put it to me). A while back I wrote a paper which contained a defense of the copying of software (and artwork, books, etc.) [Bringsjord 1989]. Four years later, I still get hate and love mail. The love mail says, "Gee, you took this copying mess and produced a formal argument which shows that copying is ethically permissible -- great!" The hate mail says, "Man alive, you shouldn’t take something like copying, which we’ve been talking about quite profitably in nice English for a long while, and go and try to formalize it and churn out a black-and-white verdict!"

>
>sh> I wish other readers would jump in here, but, on my reading,
>
>sh> what you keep doing is formalizing very simple statements until
>
>sh> one (or I, at least) can no longer keep track of the mnemonics of
>
>sh> what each means, and then you draw formal conclusions that I
>
>sh> would certainly not endorse from my original construal of the
>
>sh> statements.

The essence of error in these matters is to advocate P and to then be shown that P implies Q, where Q is false. There are countless formal consequences which follow from philosophical statements. Sometimes these consequences, though unpalatable to those uttering the statement, *are* consequences of what they say. Prolly just did it again, above, w.r.t. to logical possibility and proof, no?

>
\>sh> This would be fine if the subject matter here were MORE
>
\>sh> complicated, and the notation were to SIMPLIFY and CLARIFY it,
>
\>sh> but what I find instead here is that all the complication comes
>
\>sh> from the notation itself, and at some point the intended
>
\>sh> interpretation of the original statements is lost, and we are simply
>
\>sh> following lockstep with some algebra that no longer has any
>
\>sh> connection with what we meant.

Ah, this is wonderful! First, the purpose of the notation is to achieve a level of precision not otherwise attainable (more on that later). Second, let me show you, via what you say next, why you’re wrong, why the analysis does sometimes simplify and clarify. You say:

>
\>sh> Let me put it simply here with reference to what you said above
>
\>sh> about what might be provable a priori about Nebulons and
>
\>sh> performance capacity: I cannot imagine a formal argument you
>
\>sh> could use for the fact that if a system passed T3 but were of a
>
\>sh> particular kind (say, a finite state automaton) then it COULD NOT
>
\>sh> (with the force of either logical or physical necessity) have a mind
>
\>sh> -- for the simple reason that (for all I know, or can even conceive
>
\>sh> of knowing) even a rock or an electron could have a mind.

There’s a difference between the claim that the following two propositions are inconsistent,

(1’) It’s logically possible that P.

(2’) There exists a proof that ~P.

and the claim that these two propositions are inconsistent:

(1'') It's logically possible that P.

(2'') There exists a proof that it's not logically possible that P.

You're conflating the two claims.

Next revelation: The source of the problem (as you've no doubt realized by now), is that you assume that from (2') it follows that necessarily ~P, which is equivalent to not logically possibly P -- which would indeed contradict (1'). But now look at all the progress we've made! We now understand the source of the problem.

>sb> Look, if I'm told that your pen-pal, Pen, is a (computing?)
>sb> machine, *and* I reflect on whether or not Pen is a person, I'll
>sb> quickly come to the conclusion that Pen isn't. Why? Because in
>sb> order for Pen to be a person there must be something it's like to
>sb> be her on the inside. But there's nothing it's like to be a
>sb> computing machine on the inside. Hence Pen isn't a person.
>sb> (Spelled out in Chapter I, of *What Robots
>sb> Can & Can't Be*.)
>sb> Because in order for Pen to be a person she must have the
>sb> capacity to, within certain parameters, introspect infallibly. But no
>sb> computing machine can have such a capacity. Hence Pen isn't a
>sb> person. (Spelled out in Chapter IX, of *WRC&CB*.) Because in order
>sb> for Pen to be a person she must have free will. But no computing
>sb> machine can have free will. Hence Pen isn't a person. (Spelled out
>sb> in Chapter VIII, of *WRC&CB*.) Etc., etc.

>
>sh> Alas, I have read all these chapters, but I do NOT find them to
>
>sh> support any of the conclusions you adduce above.

Fine, but to repeat: the issue is then the truth or falsity of the premises, since the reasoning is formally valid. What's false? That's the way the game works. I give you a formally valid argument for my position. If you don't embrace my position it's incumbent upon you to say which premise is false, and why.

>sb> Now tell me what premises are false in these arguments, or tell
>sb> me to flesh out the arguments on the spot, in plain English (which
>sb> can't be done, at least not well: more about that later), but don't
>sb> tell me that which convicts you of *petitio principii*: don't tell me
>sb> that one simply can't deduce things about Pen from my
>sb> Chesterfield. Whether or not one can is precisely what's at issue!

>
>sh> Alas, I have to say exactly what you enjoin me not to: For a
>
>sh> starter, how is any formal argument going to penetrate the other-

>
>sh> mind barrier with the force of proof? (I've already agreed that
>
>sh> cheating, chance and the Chinese Room Argument against a T2-
>
>sh> passing computer have the force of a plausibility argument, but
>
>sh> PROOF?)

We've got it all figured out now. Your position stems from your private construal of 'proof.' My position stems from my public construal of 'proof.'

>sb> What I've said is that thought-experiments exploiting
>sb> serendipitous passing of TT show that certain construals of
>sb> Turing's main thesis are false, because these construals put the
>sb> thesis in the form of a conditional, and the thought-experiments
>sb> are cases in which the antecedent is true but the consequent isn't.

>
>sh> But I immediately concede that the conditional that "if X passes
>
>sh> T2 (or T3) it NECESSARILY has a mind" is false, because the
>
>sh> passing could be a result of chance, cheating, or mindless symbol-
>
>sh> manipulation (Searle). The first two of those counterexamples are
>
>sh> less an invalidation of T2/T3 as evidence than of T2/T3 as proof.
>
>sh> (Searle's argument also invalidates T2 as evidence, in the special
>
>sh> case of a [hypothetical] implementation-independent T2-passing
>
>sh> symbol system). No formalism was needed to see any of this.

No formalism was needed for *you* to see this -- wouldn't that be a bit more circumspect? And wouldn't it be true to say that you purport to see a lot in the absence of formalism? And, wouldn't it be true to say that I purport to see very little in the absence of formalism?

PART III: ...HARNAD ON THE BRINK

>sb> I think you've got things reversed. [My] Chapter... VIII's argument
>sb> from free will is a formally valid argument -- that much can be
>sb> proved. Its conclusion is that person-building AI is doomed. Since
>sb> it's formally valid, if its premises are true, person-building AI is
>sb> doomed. Your response to this is that it's too much of a thicket?
>sb> When people say that about Searle's CRA, they need to be walked
>sb> through a very carefully stated version of the argument -- a
>sb> version Searle never produced. You simply buy Searle's *idea*.

>sb> You look at Searle’s thought-experiment and you say, Yes! That’s
>sb> fortunate for Searle, but rather uncommon...

>
>sh> Searle’s argument is transparent. In stating it more carefully than
>
>sh> Searle did (as I have tried to do), one makes it even more
>
>sh> transparent. Your version, on the other hand, makes it more
>
>sh> opaque (if it’s even the same argument at all -- I can’t tell.)

>sb> HIGH-LEVEL ENCAPSULATION OF CHAPTER VIII, *WHAT ROBOTS
>sb> CAN & CAN’T BE*:

>sb> (1) If every event is causally necessitated by prior events,
>sb> then no one ever has power over any events (because,
>sb> e.g., my eating pizza was ineluctably in the cards when T.
>sb> Rex roamed the planet).

>
>sh> (I happen to think the antecedent here, and hence the consequent,
>
>sh> is true, but I certainly don’t see any reason to make anything in
>
>sh> robot-building depend on any such controversial metaphysical
>
>sh> conjectures.)

Well, we do get down to some rockbottom issues, don’t we? I think you’d be happy enough ignoring all attacks on "Strong" AI save for CRA! Look, for decades now the following kernel has been floating around: "Machines don’t have free will. They are completely deterministic. They do exactly what they’re programmed to do. But people -- ah! That’s quite a different story. People are free, they are at liberty, ceteris paribus, and within particular spheres of thought and deed, to do what they want. People have *autonomy*. They can deliberate over an issue and then make a decision -- *their* decision. And these decisions can be free of socialization, advice, etc., for the simple reason that the deliberation can take these forces into account. So, people can’t be machines." Now I have taken this kernel, this unconvincing, compressed kernel, and I’ve rendered it razor sharp -- for anyone who can come to understand the moves in Chapter VIII of *What Robots Can & Can’t Be*. And of course my argument is sensitive to the literature on determinism, free will, indeterminism, etc.

>sb> (2) If it’s not the case that every event is causally
>sb> necessitated by prior events, then, unless *people, not
>sb> events, directly bring about certain events, and bring
>sb> about their bringing about those events* (the doctrine is
>sb> known as iterative agent causation), no one ever has
>sb> power over any state of affairs (because, e.g., my eating
>sb> pizza would then simply happen uncaused and "out of the

>sb> blue," and would thereby not be anything over which I
>sb> have power).

>
>sh> This stuff is almost more controversial than the T-testing itself.
>
>sh> How can one imagine grounding CONCLUSIONS about T-testing on
>
>sh> such premises? "iterative agent causation," sui generis causality!
>
>sh> I'm just interested in whether T3 is a reliable guide for mind-
>
>sh> modelling...

The free will angle is controversial. So what? In the chapter in question I include a defense of every premise in the argument we're considering. You would no doubt say the same thing about emotions. Emotion is a controversial topic. So what? I still think it's possible to put together a formidable emotion-based attack on "Strong" AI.

>sb> (3) Either every event is causally necessitated by prior
>sb> events, or not (a tautology!).

>
>sh> Yes, but also, I think, a red herring here. ("You either have
>
>sh> stopped beating your wife, or not" is a tautology too...)

Sorry? (3) is a used premise in the top-level argument. Therefore it can't be a red herring.

>sb> Therefore:
>sb> (4) Unless iterative agent causation is true, no one ever has
>sb> power over any events.
>sb> (5) If no one ever has power over any events, then no one is
>sb> ever morally responsible for anything that happens.

>
>sh> The introduction of yet another conjecture, this time ethical
>
>sh> instead of metaphysical, and of still more controversial concepts
>
>sh> (power over events, moral responsibility)... This, together with the
>
>sh> formalism, is what I mean by making things more opaque istead
>
>sh> of clearer.

These are not conjectures. (4) follows from (1)-(3). (5) is a premise in the top-level argument, but a conclusion of a supporting argument given in the chapter! Though, (5), were it a conjecture, would be a pretty decent one. If Smith doesn't have power over anything that happens, including the knife's being driven into Jones' chest, then Smith isn't morally responsible for the murder. (5) is affirmed daily across the globe.

>sb> (6) Someone is morally responsible for something that happens.

>
>sh> Many are ready to contest this. But look, this "proof" is bringing in
>
>sh> more and more controversial stuff.

It's controversial to say that someone is morally responsible for something? I think, again, that at times like this students provide a healthy slap in the face. Students know that if their professors arbitrarily assign failing grades because they feel like it, these profs are in the wrong. And if such an F is assigned, and a student reading his report card is thereby upset, the prof is morally responsible for getting the student upset. *This* is controversial?

>sb> Therefore:
>sb> (7) It's not the case that no one ever has power over any events.
>sb> Therefore:
>sb> (8) Iterative agent causation is true.
>sb> (9) If iterative agent causation is true, then people can't be
>sb> automata (because if people can enter an infinite amount
>sb> of states in a finite time, were they automata, they would
>sb> be able to solve unsolvable computational problems).

>
>sh> Doesn't follow. They may not be able to enter the "right" states to
>
>sh> solve the mathematical problems. And "iterative agent causation,"
>
>sh> whatever its pedigree, sounds pretty dubious to me.

You're on to something here, I concede. We've have arrived at the real issue, and tough spot in the argument. I'm going to have to let *WRC&CB* speak for itself on (9).

>sb> Therefore:
>sb> (10) People aren't automata.

>
>sh> All the force of this argument was in the various premises
>
>sh> introduced. Unlike the axioms of arithmetic or geometry, these
>
>sh> premises are FULL of things to disagree with, some of them even
>
>sh> more controversial than the Turing Test itself. So how can you
>
>sh> hope to base a "proof" of the validity or invalidity of the Turing
>

>sh> Test on them?

We come full circle, and you forget where we've been, I fear. I started this whole line by pointing out that proofs in arithmetic are themselves controversial, and greatly so. To repeat, in encapsulated form: I can offer I perfectly classical proof of the proposition that Goldbach's Conjecture is either T or F -- and for principled reasons a constructivist will utterly reject the proof. A technical philosopher must come to grips with this. A technical philsopher must also come to grips with the Paradoxes -- they, too, cast a shadow of doubt across the landscape you seem to embrace unreflectively.

>
>sh> And look: From Searle's argument I have learned that people
>
>sh> aren't just implementation-independent implementations of
>
>sh> formal symbol systems. Even if I really learned from your
>
>sh> argument that people weren't "automata," what would that make
>
>sh> people instead? and what is one to donext? Searle's argument and
>
>sh> the symbol grounding problem turned us from the T2-symbolic
>
>sh> road to the T3-hybrid road. What would be the corresponding
>
>sh> guidance we would derive from this "proof" that people aren't
>
>sh> automata?

Interesting questions, all. I think we both lead relatively uncluttered intellectual lives. You've got it all cleaned up with Searle's help, and there you are walking, with a whistle and spring in your step, down the T3-hybrid road. Of course, I think I've got it all cleaned up too: I can walk down your road, *but only as a hacker.* I can work 8 hours a day on a nitty gritty effort at getting a computer to tell stories, represent informationin the FLEX logic programming environment, think about how to represent plans, actions, times, and so on -- I can do all this and more, but all the while remember with a smile that rivals yours that I'm just an ad hoc engineer. No real mentation is ever gonna come out of this. No real emotion, no qualia, no agents with free will and introspecive capacities. As I walk down the road I may not only build clever little artifacts, I may also learn a good deal about the human mind. That's good. That's what I want. But I'm quite convinced that this "Strong" AI/Cog stuff, whether or not it's got sensormotor sensitivity, is all a bunch of nonsense. People are astonishing creatures. They are capable of reasoning that is light years beyond a Turing machine, whether or not it's outfitted with sensors and effectors. I think you're going to come round to this view. Penrose doesn't establish it, but he feels it in his bones. He's probably never going to be able to defend the view at an APA meeting, but feels it -- along with many others. I think *you* realize, every once and a while, perhaps in evanescent bouts of utter doubt, that you stand, with Searle, on the brink of a maelstrom of mind so wild that it may take two or three scientific revolutions before the horizon glows with an *authentic* promise of building ourselves. [See, in connecton with my speculation, the excellent (Harnad, 1991).]

Selmer Bringsjord

REFERENCES

Bringsjord, S. (in progress) "In Defense of Searle on Zombies."

Bringsjord, S. (forthcoming-a) "Searle on the Brink," forthcoming in *Psyche*

Bringsjord, S. (forthcoming-b) "Could, How Could We Tell If, and Why ShouldQAndroids Have Inner Lives," chapter forthcoming in Android Epistemology, JAI Press, Greenwich, CT, Ken Ford & Clark Glymour, editors.

Bringsjord, S. (1992) *What Robots Can & Can't Be* (Dordrecht, The Netherlands: Kluwer).

Bringsjord, S. (1989) "In Defense of Copying," *Public Affairs Quarterly* 3.1: 1-9.

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54.

> From: Ken Easlon
>
> A virtual human-like robot in a virtual world might believe itself
> to pass T3 (or T4....Tn), if the simulation was designed to fool
> the robot into thinking it was real.

This is a paradigmatic example of getting lost in the Hermeneutic Hall of Mirrors. What's at issue here is whether there's really anyone home in an ungrounded symbol system in the first place. If there isn't, then there's no one there to "believe" anything: It's all done with mirrors.

A homology might help: Is there really any movement (or gravity) in a computer-simulated (virtual) solar system? No. It's just a bunch of squiggles and squoggles that are systematically interpretable as planets rotating around the sun.

Well it's EXACTLY the same with "beliefs" in a virtual world robot. (The only difference is that we can SEE that the virtual planets are not really moving, whereas we cannot SEE whether or not the virtual robot is believing. So in the first case we are saved from any overinterpretative tendencies by our own senses, whereas in the second we are not. Hence the hermeneutics. But we can't see gravity or quarks either, and yet we're not tempted to say that the virtual planets have them. Well, it's the same with the mental states of the virtual robot, and for roughly the same reasons. [I hope I will be spared having to do ANOTHER round on this, this time on "virtual quarks," etc.].)

> Looking at things metaphysically, the reader might be living in a
> simulated world.

Alas, this is not very good metaphysics. My world could all be a hallucination. Fine. Or my real senses could be fooled by virtual-world input. That's fine too [but now follow carefully]: Nothing, absolutely NOTHING follows from either of these two facts that justifies supposing that an ungrounded symbol system -- no matter WHAT it's systematically interpretable as being (planet, motion, robot, mind) -- is REALLY anything more than what it really is: A stationary digital computer, flip-flopping through states that are systematically interpretable as if... etc.

Welcome to the Hermeneutic Hall of Mirrors (and the Cheshire Cat's Grin).

Stevan Harnad

> From: rickert@mp.cs.niu.edu (Neil Rickert)
>
> My disagreement with Harnad is on what constitutes a computer. His
> conception of a computer seems to me to be too narrow. Some people in
> computer science and some computationalists have equally narrow
> conceptions, but other computationalists are quite broad in what they
> would consider to be part of a computational model... Searle's
> [Argument is likewise based on] overly narrow conception of computation.

Unfortunately, I inherit my conception of what a computer is from Von Neumann, Church, Turing, Goedel, and theirs too is the "narrow" view. If you mean something else by a computer than a physical system that can implement formal symbol manipulations (with the details of the physical implementation being irrelevant to the computation it is performing) then all bets are off, because we are not talking about the same things when we are saying a Computer can/can't do/be X.

> you are implying that the input-output peripherals (sensorimotor
> transducers) are irrelevant for computers. I deny this. A computer
> without suitable I/O is a piece of worthless junk. Most computer
> scientists consider at least some of the peripherals to be an
> integral part of the computer.

A computer without I/O is useless (but still a computer -- and could be chugging away at the solution of the four-color problem, but never mind). Fine. But what I said was that a digital computer's I/O is not sensorimotor transduction (and certainly not the sensorimotor transduction of a T3 robot); and that ALL the physical details of the implementation of a computation -- whether the source of its I/O or the realization of the computation itself -- are irrelevant, apart from the fact that they implement the right computation. Whatever properties a computer has purely in virtue of doing the right computation, radically different physical systems performing the same computation must have exactly the same properties. Searle showed that this fails to hold for the property of understanding (and the reason it fails to hold has NOTHING WHATSOEVER to do with the speed of computation involved -- that too is an irrelevant implementational detail); and the symbol grounding problem shows why it fails to hold.

> Once again I insist that you are misconstruing what is a computer. The
> computer as formal machine cannot do the I/O transduction. But the
> computer as physical machine with suitable peripherals quite possibly
> can. Quite obviously any such computer would look very different from
> current computers, even if only because current computers could not
> handle the volume of concurrent I/O that would be required.

By this token, an airplane is just a computer with "suitable peripherals." Fine. So by the same token, so is a T3 robot. But unhook the computer-plane from its peripherals and it is no longer a plane; and unhook the T3 robot from its T3 transducers and it is no longer a mind. In both cases, the "peripherals" are the tail that wags the dog. I see no advantage in redefining "computer" in such a wide sense as to subsume planes and robots. By the same token you could subsume furnaces, planets, etc., in which case EVERYTHING is a computer, everything is computation, and we've accordingly said absolutely nothing when we say a Computer can/can't do/be X.

Stevan Harnad

> From: rickert@mp.cs.niu.edu (Neil Rickert)
>
> A human is a physical system that can implement formal symbol
> manipulations. Somehow I don't think that was what you had in mind.

No indeed. What's at issue is the converse: Is (just) a physical system implementing (implementation-independent) formal symbol manipulations a mind? Answer: No, because of the symbol grounding problem. (I will not get into why I consider Searle's CR Argument correct; enough rounds have been done on that; the rest is between Achilles and the Tortoise.)

For interested readers: There is an ongoing "Symbol Grounding Discussion Group" whose existence predates the formation of comp.ai.philosophy. A lot of this ground has been systematically covered there, including the current issue of implementation-DEpendent "computation." Here is its InterNIC Directory of Directories entry:

Resource Name: Symbol-Grounding-List - Symbol Grounding Archive and Discussion List Resource Type: FTP Archive; Scientific Discussion Group Keywords: archive, discussion, symbol grounding, robotics computation, cognition, representation, mind, consciousness, brain, psychology, turing test, philosophy
==
Description: Discussion group and archive devoted to the symbol grounding problem. What is Computation? Is Cognition Computation? If not computation, or not just computation, then what is cognition? How can the formal symbols in a symbol system be grounded in what the symbols are ABOUT without the mediation of a human interpreter? What is the status of the Turing Test and of Robotic embedding? The Symbol Grounding Dissussion Group and Archive have been in existence and ongoing for several years. If you wish to join the discussion you must first read the archive to familiarize yourself with what has already been said. Occasionally, segments of the discussion appear in print in scholarly journals. The discussion is moderated and only serious contributions from scientists and scholars will be accepted. Access: FTP archive host - princeton.edu Directory - pub/harnad/Symground Email requests to join group or contribute discussion harnad@princeton.edu
==

> My concern is that you seem to imply that a computer can only do symbol
> manipulations. And that is just wrong. There are important applications
> for computers which are not symbolic. The programmer may still use
> symbols, but those symbols are artificial constructs. The problem of

> recording analog music, and later playing it back on an analog
> amplifier system is a completely non-symbolic task, yet it is handled
> by computers in the digital recorders and CD players. If you presume
> that computers are limited to symbolic problems, you misunderstand
> their capabilities.

The computer encodes the sound, and with suitable acoustic transducers, can mediate both its recording and playback. There is no "music" in the computer itself (any more than there is any in a static musical score on paper, or even on a disk or tape). The dedicated connections of the computer's symbolic states to the acoustic transducers "fix" the interpretations of its symbols in a way that is analogous to how a robot's capacity for T3 interaction with the objects its symbols are about grounds its symbols. But without the T3 capacity we are back to a digital computer that no more thinks than it sings.

And I don't for a moment think the sensorimotor transduction in the T3 robot's case will be commensurable with the simplicity of acoustic transduction for audio recording/playback; nor do I believe that computation will be doing a significant part of the footwork in the robot's case. But that's just quibbling over details. The upshot is that a computer MINUS its audio, aero or sensorimotor peripherals is in precisely the same boat with respect to making music, flying, or thinking -- it's just the UNOBSERVABILITY of the last of these three [because of the other-minds problem, which is absolutely unique to cognition] that leaves us so much room to project our hermeneutic fanstasies unchecked.

(By the way, at the binary level, it's all symbolic, whether the higher-level language is linguistic, logical, numerical, or what have you. More important, the algorithms are just syntactic rules operating on the [arbitrary] SHAPES of those symbols. Von Neumann would agree with Turing, Church and Goedel on that.)

> The speed of the computation is not irrelevant. By my estimation, if
> Searle spent 100 years operating the Chinese Room, he could perform the
> equivalent of only about 1 millisecond of computation for the supposed
> AI system he is emulating. Now 1 millisecond of activity is probably
> below the threshold of human awareness, so if the AI system is a proper
> implementation of a mind we would not expect any consciousness of that
> 1 millisecond of activity. Thus it is not at all surprising that Searle
> does not understand. Indeed, it is predictable from a strong AI
> assumption.

I can only repeat, there is no principled reason, either from the theory of computation or (what there is of) the theory of cognition, for even faintly asserting that the (normally irrelevant) implementational details (whether temporal, spatial, or what have you) of the physical implementation of any computation should have any bearing whatever on whether it is implementing X -- IF the thesis in question is "X is Just a Form of Computation" (as it is in the case of Strong Computationalism about Cognition). The temporal numerology is pure ad hoccery in the service of theory-saving in this context. You should just be aware that if you are putting your theory-saving money on implementation-DEpendent "computation," then you are NOT saving the thesis of Strong Computationalism but changing (or clouding) the subject (to sci-fi speculations about what is special about silicon).

> The airplane is not autonomous -- it requires a human pilot. If you
> take a pilotless aircraft, such as a cruise missile, it does not sound
> nearly so strange to refer to it as a flying computer.

But it sounds even less strange to refer to it as merely a plane -- with it's flight guided by a computer...

Stevan Harnad

----

ftp host: princeton.edu directory: pub/harnad/Harnad

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25. FILENAME: harnad89.searle

Harnad, S. (1990) The Symbol Grounding Problem. Physica D 42: 335-346. FILENAME: harnad90.sgproblem

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on Virtual Mind. Minds and Machines 2: 217-238. FILENAME: harnad92.virtualmind

Harnad, S. (1993) Grounding Symbols in the Analog World with Neural Nets. Think 2(1) 12 - 78 (Special issue on "Connectionism versus Symbolism," D.M.W. Powers & P.A. Flach, eds.), followed by 13 commentaries (including Searle's) plus responses. FILENAME: harnad93.symb.anal.net

----

> From: rickert@mp.cs.niu.edu (Neil Rickert)
> Date: Tue, 1 Feb 1994 02:07:02 GMT
>
> [I] magine that I can get a complete map of your neural system. I will
> attach a symbolic label to each neuron, and attach a symbol to each
> neural signal. Suddenly, and with no physical change whatsoever to your
> body, you (or your neural system) have become (just) a physical system
> implementing formal symbol manipulations. If your claim is correct, my
> act of labelling has destroyed your mind, even though it made no
> physical changes to your physical body.

Here's a useful criterion for distinguishing things that are really X from things that are merely interpretable as if they were X (whether X is a brain, a plane, a concert performance, or a mind): EVERYTHING that is X is also (trivially) interpretable as if it were X. But only things that are really X are MORE than merely interpretable as if they were X.

When you do a labelling like the one you are proposing, you are just showing that the brain (like everything else -- except perhaps truly continuous systems and perhaps turbulent ones) is computationally equivalent to a computer running some program. That's just (the physical version of) the Church-Turing Thesis (to which I subscribe, by the way!).

So let’s accordingly note that for every real thing, there is also (at least) one virtual counterpart. Since that virtual counterpart is implementation-independent, it can be a digital computer simulation, or even the thing itself, suitably "labelled." So what? The difference between a real and a virtual plane is still quite clear: BOTH of them are interpretable as if they were planes, flying (they can even be so labelled), but only one of them is REALLY really flying, hence really a plane. The other is JUST an ungrounded [or should I say unflying] symbol system that is merely interpretable as if it were a plane flying (with its implementation irrelevant, whereas that of the real plane is decidedly relevant).

Precisely the same is true of a real brain, a "labelled" real brain, and a virtual brain. The first two are the same thing (and interpretation-independently so), whereas the last is something else -- and has, in a paradoxical way, more in common with the virtual plane than the real brain: They are both just symbol systems that are systematically interpretable as if they were something else; what they are is completely interpretation-dependent (actually, it’s a bit more powerful than that: it’s interpretability-dependent, but let’s not get into that; nothing hinges on it).

That, again, is the symbol grounding problem. It is easy to break out of the hermeneutic circle in the case of a plane: What you need is something that actually flies, and is not just symbol-crunching that is interpretable as if it were flying. I’m proposing that you need is something rather similar in the case of the mind, thinking, except that because thinking, unlike flying, is unobservable, you can never have the same certainty that you have in the case of observables. Yet the constraint I propose is observable too (it’s just that what’s observed is unfortunately not thought): The candidate must have robotic capacity that is T3-Indistinguishable from that of people with minds.

The fact that a digital computer (or any implementation-independent implementation of any symbol system, no matter how it can be interpreted) cannot be that winning candidate has nothing at all to do with whether or not you can either label or simulate brains. You CAN label them, but only because you can label and simulate just about anything (the Church-Turing Thesis is hence really just Weak Computationalism). Yet thinking is not the kind of thing that an ungrounded symbol system can do, any more than it can fly, and for very similar reasons.

> In that case [if there is no music in a computer, disk, or music score]
> there is not any music in your ears either. This all depends on how you
> define music. Let me make it clear that I was not in any way suggesting
> that a CD player has a mind.

I didn’t think you were, but none of this has anything to do with definitions. Music is an ACOUSTIC phenomenon, not an implementation-independent computational one. Musical experience, in turn, is an AUDITORY phenomenon, normally caused by acoustic stimulation, but capable, just like thinking, of occurring without acoustic stimulation (and in the absence of any observable counterpart we can be sure of). In a digital computer (or disk, or musical score) there are neither acoustic nor auditory phenomena going on (though, again, because the latter are unobservable, hermeneutics again beckons). In a digital computer there is only ungrounded symbol crunching going on. That is not true of what is going on between my ears. Therefore there must be more to what’s going on between my ears than what is going on in an ungrounded symbol cruncher. Whatever that is (whether it’s something like my proposed hybrid analog/neural-net-symbolic model or something else), it’s not JUST implementation-independent symbol manipulation that is systematically interpretable as if it were auditory (or acoustic).

> Clearly temporal details are not irrelevant. If you were designing a
> robot, whether TTT or anything else, you would discover that timing can
> be critical. As for the thesis "X is Just a Form of Computation", the
> validity of this is highly sensitive to the definition of X and the
> definition of computation.

To the extent that speed of computation (as opposed to speed of T3 behavior) is relevant to generating T3, thinking is not just computational. I agree that this follows from the definition of COMPUTATION (in which speed does not enter, because computation is implementation-independent), but it certainly does not follow from the definition of COGNITION, since no one knows what thinking is, we just know we do it, without the faintest idea how. (One is ill-equipped to rely on definitions under such conditions!) One thing that I suspect is true of cognition, no matter what it turns out to be, however, is that it is more like flying than like implementation-independent symbol manipulation; and that whether or not something is thinking is more than just a matter of interpretation (or interpretability). At least that’s true of MY thoughts, which mean what they mean independently of what they are interpretable or interpreted by anyone else as meaning.

> Let me give you an example of implementation dependence which I suspect
> is probably important. If I am designing a robot, the internal logic
> must keep track of the position of the robot arms as they are moved.
>
> Method 1: (The method of the frame problem) Compute the new
> position of the arms for each motion, and store the
> computed position in internal data structures.
>
> Method 2: (The method of perception) Use visual input from
> the robot eyes and data from proprioceptive sensors at
> the various joints to determine the position of the
> arms, and update the visual and proprioceptive data
> continuously.
>
> Now I happen to think that method 2 is the correct method. You seem to
> be implying that it does not matter which of the methods is used.

I’ll bet on whatever scales up to T3 (I think that will narrow the degrees of freedom to the same size as the ones facing the Blind Watchmaker, and that’s close enough for me), and at that scale, yes, I would say the differences make no difference (or, if they do, we can never hope to be any the wiser). Whether a toy task is accomplished this way or that is of no interest whatever (because a toy task is hopelessly underdetermined): only what will successfully scale to T3 is of interest. Now I happen to be betting on (shall we call it) "analog-intensive" hybrid models more like 2 than like 1, above, but only because I have repaon to believe they have a better chance of scaling. I’m already covered by the essential role of the sensorimotor transducer component in implementing T3 power (and hence mind) either way.

Harnad, S. (1992) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. SIGART Bulletin 3(4) (October) 9 - 10.

Harnad, S. (1993) Problems, Problems: The Frame Problem as a Symptom of the Symbol Grounding Problem. PSYCOLOQUY 4(34) frame-problem.11.

Stevan Harnad

> From: pindor@gpu.utcc.utoronto.ca (Andrzej Pindor)
> Date: Tue, 1 Feb 1994 22:02:09 GMT
>
> If someone becomes paralyzed, loses, sight, hearing, etc., does he/she
> lose the grounding too? The correlations referred to above are merely
> ENCODED in his/her brain.

COMPUTATIONAL HOMUNCULARISM AND UNGROUNDED CORRELATION

There is a fallacy that courses through Pindor's Reply here, and through much of the discussion surrounding it, and I think "computational homuncularism" is probably as good a name for it as any. The fallacy is thinking that mental states, at bottom, HAVE to correspond to the activity of an internal module (the homunculus), which is very much like a computer, with all the rest just its peripherals and data. Being in the grip of this homuncular view, it is very hard to imagine that WE might not be such computational modules, but rather the "peripherals" themselves.

The hypothesis that is ON TRIAL here ("computationalism") is that the brain is like a digital computer executing the right program: that it is just an implementation-independent implementation of the right symbol system (the T2-passing one, in most examples). The symbol grounding problem (and the Chinese Room Argument) imply that this hypothesis is false. One cannot, accordingly, ANSWER the argument or solve the problem by simply coming back again as if the brain WERE like a digital computer running the right program after all.

In a nutshell: BECAUSE of the symbol grounding problem, there are strong reasons for rejecting the idea that a person who is paralyzed, blind, etc. is like a computer with its peripherals removed. [Moreover, the neurobiological evidence suggests that what a brain with ALL of its sensorimotor transducers (all its afferent and efferent connections and activity) ablated would be would be DEAD (or just a hapless hunk of protoplasm) rather than a computer running a program that implements a solipsistic mind.]

BUT (and this too follows from the symbol grounding problem and Searle's Chinese Room Argument): IF the completely de-afferented, de-efferented brain WERE just reduced to a computer running the "right" program, then it would follow from symbol grounding considerations (and the CR) that it WOULD (like any other implementation-independent implementation of a symbol system) thereby lose its grounding, and hence would not have a mind (unless there was something special, i.e., noncomputational, about that particular implementation -- which would still not be good news for the hypothesis on trial here).

And, conversely, if the de-afferented, de-efferented brain DID still have a mind, all that would follow from that would be that it was NOT just a computer computing (i.e., implementation-independently implementing the right symbol system). It must have other properties, noncomputational (i.e., implementation DEpendent) physical/physiological properties, and THOSE would be the properties that ground its symbols -- and that WOULD generate T3 capacity, if only its afferents and efferents were restored. I am using "sensorimotor transduction" as a stand-in for any and all of the noncomputational, implementation-DEpendent properties that a hybrid system would have to have in order to be grounded. But no matter what point you single out to pull the plug on the grounding, the argument is that there will be no mind left on the other side. Imputing a computational homunculus to whatever is left over is simply theory saving (or wishful thinking). At some point one has to concede that there may be good reasons for thinking that it’s NOT all just input to a computer with a mind; what one had thought was just input or peripherals may just be an essential part of the homunculus (who is then no longer a computational homunculus, getting data, but a much more hybrid entity corresponding to US).

Speaking about correlations being "encoded" in the brain is completely equivocal, because it is the interpretation of the codes that is ungrounded. Thoughts are ABOUT things, and if the connection between a state and the thing it is allegedly about must be mediated by an external interpretation, then that state is not a thought. State-internal correlations, no matter what they "encode" ("represent," "mean," etc.) will not get you out of this symbolic circle (the hermeneutic circle). That IS the symbol grounding problem.

> [Concerning the possibility that Searle could somehow learn Chinese by
> manipulating Chinese Symbols in the Chinese Room]: Please note what you
> have written yourself in [Harnad1993] about a lack of proof for
> existence of semantic duals and in [Harnad1990] about decrypting
> ancient languages. In light of this Searle could very well come to
> understand Chinese (Neil Rickert remark about speed of processing is
> very relevant here)

Sure Searle might decrypt the Chinese symbols eventually. So what? We KNOW Searle has a mind. But he was supposed to be FURTHER implementing something that IS a (Chinese-understanding) mind, by implementing the same program the T2-passing computer implements. Does the computer also decrypt its symbols eventually? (The infinite regresses we encounter at every turn are the characteristic symptoms of the symbol grounding problem.) Searle’s possible eventual decryption of the symbols is completely irrelevant to the Chinese Room Argument; it is merely an example of what external interpreters with grounded minds can do with ungrounded symbol systems (which is just as irrelevant as speed of processing, an implementational detail that can have no bearing on the hypothesis that cognition is just computation).

> [Concerning the possibility that manipulating Chinese
> Symbols in the Chinese Room could somehow generate a second,
> Chinese-understanding mind in Searle of which he was unaware]: This is
> a possibility a lot of people accept, and you cannot rule it out.
> Searle himself cannot know this, since he is NOT this second mind. The
> second mind is what converses with you in Chinese. If you think that
> this is ridiculous (that Searle might be unaware of it), please note
> that the notion of Searle being able to squeeze into his mind all these
> rules and data required for English-Chinese translation is no less
> ridiculous. Why do you accept one and refuse to accept the second?

The second merely requires me to suppose that Searle would just have to do more and more of the same (with no inductive reason whatever to suppose that some sort of "phase transition" into mental space would ensue if he could just do enough of it, fast enough). This is simple induction, requiring no stretch of the imagination or special pleading.

The first, in contrast, requires me to believe that merely manipulating a bunch of meaningless squiggles and squoggles would generate another mind in a person, one he wasn't aware of -- an outcome that normally requires something as radical as childhood sexual abuse to induce. Why on earth should anyone believe something like this could be caused by mere symbol manipulation?

The (huge) difference between the two is merely one of plausibility. Possibility/impossibility was never the issue (how could it be, with the other-minds problem always there as a spoiler)?

> The only thing which is necessary is that the symbols for mushrooms and
> toadstools be _different_. You have nowhere established that their
> actual shapes count.

Necessary for what? I'm talking about actually sorting and labelling mushrooms. To be able to connect the right symbol with the right object, the system has to pick out the invariant properties of the sensory projection that allow it to perform the categorization correctly (T3-scale). The shape of those projections and of those invariants is nonarbitrary (and nonsymbolic). In my model, they (and the neural nets connecting them to their respective symbols) are PART of the hybrid system that implements mental states. Ablate the nonarbitrary shapes and the system loses its T3 capacity, hence its grounding, and hence (by my lights) its mind. Substituting your computational homunculus for my hybrid analog/connectionist/symbolic system is simply regressing to the fallacy I mentioned at the outset (and changing the subject, which in the categorization work was sorting real mushrooms, not performing internal correlations).

> You can conceivably imagine an advanced virtual reality set up where
> senses are bypassed - computer generates necessary electrical signals
> which are fed directly to optic nerve. We simply skip creating an image
> which eyes then convert back into electrical signals. This happens
> externally to the brain (mind) and I do not see how it can influence
> it. Now we have NO transducers - your argument falls down.

The optic nerve is still a peripheral ganglion. If you stimulate a nerve ending, you are still doing transduction, this time of energy on the nerve membrane instead of the sensory sheet. So what? Even if you went straight to the visual cortex it would be transduction, because you've got to get that stimulation "into" the brain. And the brain is not just a digital computer, getting data. Our real experience of virtual reality is generated by stimulating our real senses (which stretch pretty deeply and pervasively into the brain, by the way). The very same computer-generated input sent directly TO a computer instead of human senses would not be an experience for anybody. Now why is that? (This version of the symbol grounding problem is what I called the Cheshire Cat's Grin...)

> Cryptologists (as you have noted yourself in another place) break out
> of [the dictionary] "go-around".

Yes, but cryptologists have MINDS. You can't project THAT a priori onto a computer without falling into an infinite regress.

> If the structure of correlations
> underlying the symbols maps sufficiently well outside world
> relationships, the symbols should be considered grounded.
> That is all you can ask for.

I certainly CAN ask for more: "Considered" grounded, systematically interpretable as grounded, Turing-equivalent to grounded -- none of these are what's meant by GROUNDED, but rather the opposite. in fact, "considered grounded" is virtually an oxymoron in this context. Symbols have to BE grounded, independently of external interpreters' "considerations." What my thoughts are and are not about certainly doesn't depend on such considerations. How to get MORE than a system whose symbols can be systematically CONSIDERED grounded IS the symbol grounding problem. Correlations (between symbols within the system or between system-internal states and external objects or states) are not sufficient for grounding. What is needed is an autonomous, interpreter-independent CAUSAL CONNECTION between symbols and what they are about. That's what T3 and sensorimotor transduction provide.

> True, achieving such mapping may require full robotic
> capability and certainly this structure of correlations is most
> directly obtained using the full robotic capability. Also, effective
> interaction with the outside world is the best proof that the symbols
> of the system provide such mapping. However, if I become paralyzed and
> loose all my senses, do you think my symbols immediately loose their
> grounding? Few people would agree with this and I am sure this would be
> an incorrect way of looking at grounding.

Again: the homuncular fallacy that robotic interactions are just INPUTS to the homunculus, or ways of testing whether anyone's home, and that at the limits of sensorimotor denervation would be that computational core that was ME. Consider that both you and your grounding may be in the many layers of brain you are imagining peeling off here...

> Well, I do not know about you (:-)) but I am NOT a transducer. As I
> have said above, even if I lost all my transducers, I would still be ME
> (at least for some time; impermanence of memories is another issue).

You seem pretty sure of that; the only thing *I'm* that sure of is that my thoughts don't mean what they mean merely because they can be so interpreted by you. I have DEDUCED from that that I must be more like a (sensorimotor) transducer than an implementation-independent implementation of a symbol system. You, on the other hand, seem to be satisfied with systematically interpretable correlations.

> Your pump does not work for me. Note also a possibility of the advanced
> Virtual World (no transducers, required electrical signals fed directly
> to sensory pathways) I have mentioned above, i.e. brain-in-the-vat.

And your computational-homunculus-in-a-vat unfortunately does not work for me.

>
\>sh> You could build even a T3 robot without any real-time history of any
>
\>sh> kind, and its internal symbols would be grounded
>
\>sh> (interpreter-independent) in virtue of its T3 (CAUSAL) capacity.
>
> You mean right off the bat? I do not think so. A lot of grounding is
> acquired in the process of early brain development, when proper
> structures are formed providing correlations among various sensorimotor
> modes. T3 capacity is not enough by itself.

You are merely describing the way a brain normally reaches its mature state, not the necessity of the real-time history. Besides, children (and animals) have minds, so their internal symbols (if any) are grounded. (They also have their respective T3 capacities.)

> Then you are claiming that a person who lost its sensorimotor capacity,
> becomes immediately ungrounded? What about a person in a
> sensory-deprivation tank?

There are no people who are COMPLETELY de-afferented and de-efferented. If they were, bodily function would collapse. (And sensory deprivation, like virtual reality, has nothing to do with any of this.)

>
\>sh> To put it another way: A T3 system, sitting here, idling, is still
>
\>sh> perfectly grounded, in virtue of its REAL sensorimotor capacity, a kind
>
\>sh> of physical potential it really has.
>
> I am sorry, this does not make sense - say we have a robot constructed
> in such a way that we can connect or disconnect its transducers at
> will, with one flick of a switch. Are you saying that grounding will
> depend on the position of the switch??

Sure, no problem (except that I am not imagining the T3 robot as a homuncular computational core plus trivial transducers, as you are; I am imagining it as mostly transducers, so that when you turn those off, you’re turning of most of the robot -- but never mind: my answer is yes).

> Thus, in effect, you are claiming that there can be no ’understanding’
> of abstract mathematics - if mathematical terms we are manipulating are
> not, one way or another, connected to symbols built from sensorimotor
> experiences, we manipulate them "mechanically, without understanding
> what we are doing". It is at least a controversial view.

Yes.

>
\>sh> in Searle's Chinese Room, and there really IS no understanding or
>
\>sh> meaning going on. (I doubt that that's what the formalists had in
>
\>sh> mind; certainly Roger Penrose does not think that's all that's
>
\>sh> going on in the mathematician's mind.) >
> Our symbols acquired their grounding through sensorimotor capacities,
> but they will not loose it immediately, would this capacity be lost. Or
> do you maintain that they would?

I don't think we are a computational homuncular core, so I think losing sensorimotor capacities amounts to losing most of our brains. Moreover, "sensorimotor transduction" is just my shorthand for all the analog, nonsymbolic (i.e., non-implementation-independent) components of the hybrid system I think we are. Yes, if they were lost, and we were reduced to a computational core, we'd be gone and whatever symbols were left would be as ungrounded as the ones on this screen.

> What a digital computer REALLY is depends on how it is used. Just like
> a piece of rock may be weapon if it used as such. Is it REALLY a
> weapon? Digital computer is REALLY just a bunch of atoms, for a
> Cro-Magnon man it might be an awkward piece of dead-weight.

But what *I* am does not depend on how I am used by others. So I am not just a computer.

Stevan Harnad

-------------------

> From: rickert@mp.cs.niu.edu (Neil Rickert)
> >
\>sh> only things that are really X are MORE than merely interpretable as >
\>sh> if they were X.
>
> That "definition" is nothing but question begging.
> In this case we were talking about being a symbol processing system.

HUMPTY-DUMPTY SEMANTICS

No question-begging at all. And the criterion works just as well if X is a plane, a brain, a mind, or a computer.

A digital computer is an approximation to a Universal Turing Machine. That means that it can be reconfigured by its software to become an implementation of just about any symbol system. Now just about ANY and EVERY physical system (e.g., a plane), besides being whatever it really is, is also interpretable as being a symbol system, but not as being ANY and EVERY symbol system. And when a digital computer is reconfigured to implement any particular symbol system (say, the one a plane is interpretable as being), the only property of the simulated physical system it is implementing is its systematic INTERPRETABILITY as a symbol system. But since the implementation of symbol systems is irrelevant (whereas the implementation of, say, planes, is NOT irrelevant -- if you have doubts, try flying to a real place with a virtual plane), the reconfigured computer has only one of the two properties that a real plane has.

So I repeat: only things that are really X are MORE than merely interpretable as if they were X.

What is really behind this discussion is "Strong Computationalism," which is the thesis that mental states are JUST computational states. Now computational states are implementation-independent implementations of symbolic states. So to be JUST (an implementation of a) computational state would require that being a mind, unlike being a plane, is something that just reconfiguring a digital computer into the right symbolic state would accomplish. Why would we believe this of mental states and not aerodynamic states? For the simple reason that, because of the other-minds problem, thinking, unlike flying, is unobservable by any physical means (except BEING the thinker). It is on this built-in equivocation that Strong Computationalism feeds (aided and abetted by the human mind's extreme susceptibility to be seduced by systematic interpretations, otherwise known as hermeneutics).

Ceterum sentio, I can assure you that, at least in my case, being/having a mind is MORE then merely being interpretable as if I were/had a mind.

> I have a much simpler criterion for you -- NOTHING in this world really
> is a symbol processing system. All that is possible is for some systems
> to be interpreted as symbol processing systems. The idea that any
> system, including a computer, is inherently a symbol processing system,
> is nothing more than a confusion. If I correctly interpret him, even
> John Searle would agree with this.

The Symbol Grounding Discussion Group has done many rounds both on the question "What Is Computation?" and on the question "Is Cognition [Just a Form of] Computation?" John Searle's views about what a computer is can be construed in different ways: If taken to mean that what symbol system a physical system implements requires an interpretation, that is perfectly true. But if taken to mean that anything and everything can be interpreted as being any and every computer, then I think that's either incorrect or trivial (trivial in the sense that the "interpretation" would have to be of orders of magnitude of complexity greater than the thing itself), for cryptological reasons. A digital computer can be physically reconfigured by its software to become the implementation of just about any symbol system, in other words, the computer's states are potentially interpretable as (i.e., they can simulate) any other system. This is a special property of physical systems that can be reconfigured into any symbol system. It is also the property capitalized on by Weak Computationalism (which is really just a variant of the physical version of the Church-Turing Thesis). Not every system has that property; only computers do.

And a lucky thing they do, too. Because if it WERE true that everything is a computer, then Strong Computationalism would be vacuously true (because the brain is a thing too), and completely uninformative. As it happens, it is not vacuously true, indeed it is false. (If there were no such thing as being just an implementation-independent implementation of a specific symbol system (or if being one were something one could decree Humpty-Dumpty style, of anything and everything, just by willing it to be interpretable that way), then the Chinese Room Argument itself would be incoherent; it is not incoherent, however, because there ARE multiple implementation-independent implementations of specific symbol systems, and digital computers ARE physical devices with the property that they can be readily reconfigured into just about any symbol system -- hence they are approximations to an implementation of a Universal Turing machine. So Searle can go on to use that property to become an implementation of one of those symbol systems himself [the Chinese T2-passing one] and can then tell us what the computer cannot tell us: that though his symbols are systematically interpretatable as if he understood Chinese, he in fact does not understand Chinese!)

> any claim that a symbol processing system
> cannot have a mind is a completely empty claim. Being a symbol processing
> system is purely an interpretation, and has no necessary relation to
> whether a system can have a mind.

Being SOME PARTICULAR symbol system is certainly mediated by an interpretation (which is not the same as saying it is a matter of opinion), but figuring out WHICH symbol system takes some more work (for cryptological reasons [Harnad 1993], and because not everything is every symbol system). But being ONLY that symbol system and nothing more, such that any and every implementation of it will be the same kind of thing, and will have the same essential properties -- that's something that may be true of say, a desk calculator, but, it's not true of a mind.

> That the digital computer is an implementation-independent implementation
> of a symbol system is purely an interpretation -- a very useful
> interpretation, I agree, but nevertheless an interpretation. In reality
> it is a causal electro-mechanical system. Whether it is possible for
> such a computer to have a mind is still an open question, and is not
> something to be dismissed with sleight of hand.

That a digital computer can be reconfigured by its software to be an implementation-independent implementation of (just about) any and every symbol system is a physical property of digital computers (and not, say, planes, or even brains). Whether one of those reconfigurations becomes a mind is something that is shown to be unlikely not by sleight of hand but by a closer look at the difference between being an X and merely being interpretable as being an X.

>
>sh> the definition of COMPUTATION (in which speed does not enter, because >
>sh> computation is implementation-independent), >
> One must distinguish between computation in the abstract sense, and
> computation in a practical sense on causal physical hardware. Computation
> in the abstract sense is not time dependent. But practical computation
> is always time dependent. Unless the time for completion of the
> computation is shorter than the lifetime of the computing machinery,
> the computation cannot be completed. In the Chinese Room scenario, the
> lifetime of the machinery is limited by the life expectancy of John
> Searle, the operator of that room.

To put it even more simply, computation is time-INdependent; the physical implementation of computation is NOT time-INdependent. But since the physical implementation of computation is irrelevant (to any purely computational property), its time-dependence is also irrelevant. Searle's finite lifetime and slow motor capacities are as irrelevant to the question of whether Cognition is

Just Computation as they are to the question of whether or not there occurs a string of seven consecutive sevens in the decimal expansion of pi. (Life is short, but Platonic Verities can afford to take their time.)

Harnad, S. (1993) The Origin of Words: A Psychophysical Hypothesis In Durham, W & Velichkovsky B (Eds.) "Naturally Human: Origins and Destiny of Language." Muenster: Nodus Pub.

Stevan Harnad

> From: throopw%sheol@concert.net (Wayne Throop)
>
> Assuming for a moment that regular
> analog recording equipment (say, tape) involves "transduction", what is
> it about the digital case that makes it "symbol manipulation"?
> Or am I wrong about normal recording being transduction?

Yes of course the transduction of the acoustic oscillations into anything else is transduction. (Transduction is any physical process in which energy is transformed from one form to another.) What is not transduction is its endproduct: If I transduce sound into a pattern of grooves on a record, the record has no sound, it merely ENCODES sound. The same is true if I transduce sound into dynamic or static patterns of states in a digital computer. The latter merely encodes sound. To have sound again, you again need a transducer. So whether recording FROM sound, or reproducing TO sound, the transduction is the process of turning sound into (or out of) something, not the encoding of it in between. This is exactly parallel to the T3 robot case, where I keep reminding everyone that a digital computer alone is ungrounded; only a robot is grounded, and a digital computer is not a robot. (A robot might be a computer PLUS sensorimotor transducers, but not a computer alone; that's why I keep repeating that "I am [mostly] a [sensorimotor] transducer.")

But that's not all; the T3 criterion is doing work in my theory. It is narrowing degrees of freedom, reducing underdetermination. A toy model (or subscale robot) can be constructed in a very large number of ways, most or all of them arbitrary and having nothing to do with the way we are constructed, hence there is no reason at all to think they have minds. What is it about the way we are constructed that makes us have minds? I'm betting that it's whatever it is that it takes to generate our T3 capacities. "T" stands for "Total" (just as much as it stands for "Turing": T3 is the Total Turing Test). I have dubbed this the "Convergence Argument" for T3.

So, whereas a dedicated computer plus audio equipment for recording and generating music might be called a "grounded system," the grounding is trivial (at least from my point of view, which is that of mind-modelling). Only T3-scale grounding is relevant if what we are trying to construct is something with a mind. But I repeat, take away the T3 capacity, which depends ESSENTIALLY on transduction capacity, and you have also taken away the grounding (and hence the mind, if I'm right), and all you have left, again, is an ungrounded symbol system.

Now here comes the subtlest part (but no less concrete or potentially valid for being subtle). It is subtle because, as with would-be perpetual motion machines, this is where the hermeneutic self-delusion always creeps in, and we just fall into an empty self-fulfilling prophecy: WE DO NOT KNOW WHAT A BRAIN IS OR DOES. Hence when we contemplate

Helen-Keller/Stephen-Hawking, a real person (but one who, because of disabilities, cannot pass many components of T3) we are not in the least entitled to suppose that they are just digital computers deprived of their peripherals! That they are highly unlikely to be that is precisely the thrust of the symbol grounding problem! We KNOW they have minds, but we don't know what their brain function consists of. What follows from the symbol grounding problem, on the contrary, is that there is one alternative we can safely eliminate about what is really going on in H-K/S-H heads: It's NOT just computation (i.e., not just an implementation-independent implementation of a symbol system, as in a digital computer).

It does not follow from any of this, however, that then what we must do to build a mind is to slavishly ape every property of the brain. (I advocate a form of functionalism too, but it is not the usual "Symbolic (T2) Functionalism" also known as "Strong Computaionalism"; rather, it is what I've dubbed "Robotic (T3) Functionalism". [Aping the brain, by the way, would be T4, "Neuromolecular Functionalism," which I reject as overdetermined and unnecessary.]). I am prepared to believe that T3 narrows down the alternatives just right, so that it only lets systems with minds through. But for that, the system really has to be T3! Remember that it is not H-K/S-H that is on trial here; we didn't BUILD them, and we have biological reasons for feeling as confident that they have minds as we are that the rest of us do. But when it comes to artificial candidates, the criterion becomes much more critical, and in my writings I've adduced many reasons why I am going with T3.

With this as background, it will be easier to answer the rest of the comments below (for details, see the citations at the end).

> My own perspective is, the "music encoding" of either one is
> "grounded", since both are causally connected to the actual
> sounds of the music themselves.

As I've said, a dedicated, computerized recording studio is "grounded," but only in a trivial sense, insofar as mind-modelling is concerned. But even there, neither the disks nor the computers themselves, in isolation, are grounded.

> Note: that doesn't imply the music encodings encompass *intelligence*
> or *consciousness*. Those are separate problems from *groundedness*,
> I hope it is agreed.

Agreed indeed -- but they are also the nontrivial aspect of the symbol grounding problem, since thinking (unlike sound) is unobservable, so we could be seriously wrong in attributing it to a system. Hence my T3 criterion: to reduce the underdetermination to the right size.

> the distinction drawn between "transduction" and "symbol manipulation"
> seems artificial, stilted, and leads to some odd conclusions indeed,
> which aren't (as far as I know, convincingly to me) addressed by Harnad.

The distinction may be artificial (it's certainly formal) (I'm not sure why you say it's "stilted": stilt+ed (‘stIltId) adj. 1. (of speech, writing, etc.) formal, pompous, or bombastic. 2. not flowing continuously or naturally: stilted conversation. 3. Architect. (of an arch) having vertical piers between the impost and the springing.) and it certainly leads to odd conclusions (such as the Church-Turing Thesis), but I think most of those conclusions are correct. How the theory of COMPUTATION slipped into the theory of COGNITION, however, is not only odd, but rather arbitrary and fanciful (motivated, as I have repeatedly point out, not just by the success of computers -- that already followed from the C-T Thesis -- but by the unobservability of thought and, apparently, the irresistibility of mentalistic hermeneutics).

The gist of the theory of computation, however, has nothing whatsoever to do with the mind; but it does separate, once and for all, transduction and computation:

Computation is implementation-independent manipulation of symbols purely on the basis of their (arbitrary) shapes. A computation must be implemented SOMEHOW, to be sure, but the physical details of how it is implemented are irrelevant to the actualy computation performed. And EVERY physical implementation of that same formal symbol system will be performing the same computation, so whatever (purely computational) properties it has, all implementations will have, and whatever it lacks, all implementations will lack (at least in virtue of the COMPUTATION they are performing -- they may have properties in virtue of how the computation is physically implemented, but then those are not COMPUTATIONAL properties).

So every (dynamic) physical system is some kind of transducer, and that includes digital computers. But if the system is considered AS a computer, its transducer properties are irrelevant (except for the fact that they must be implementing the right computation). So, as artificial and "odd" as it may seem. the transducer/symbol-system distinction is just the hardware/software distinction, and the implementation-independence of the software level of function.

A sensorimotor transducer AS a sensorimotor transducer, in contrast, is NOT the implementation-independent implementation of a computation. It is what it is: a physical system that transforms one form of energy into another.

> The problem I haven’t seen adequately addressed here is the
> "T3-scale".... The problem becomes
> concrete when talking about the differences between "computers" and
> "robots", and why both scale and shape aren’t significant with
> respect to humans, but *are* with respect to computers/robots.

Scale and shape are significant in both cases, but humans are not on trial here (we know they have minds and we don’t know what’s going on inside them); it is computers (and any implementation-independent implementations of symbol systems) that are on trial. So H-K/S-H need not actually pass T3, but a robot must. (Once it does pass T3, THEN we can worry about how to scale it back down to model H-K/S-H.) The purpose of T3-scale is to converge on the same functional constraints that shaped our own systems and thereby to screen out imposters. And shape is relevant for robots because the shapes of the shadows projected on them by objects are not arbitrary (whereas the shapes of symbol tokens in a symbol-cruncher ARE arbitrary, because of the implementation-independence of computation).

> exactly what is it that Hawking can still do that a computer with a
> video feed and one or two solenoid outputs cannot do? Why does anyone
> care about human shape and/or sensory bandwidth, and just how does this
> affect the grounded-ness of a computer vs that of a robot?

I’ve answered this. Sub-T3-scale toys, even "grounded" ones, prove nothing. They are hopelessly underdetermined. There are countless ways to skin a toy cat. An analogy I’ve used before is that if DC-11 aircraft grew on trees, and we had no idea how they were built or worked, it wouldn’t advance our understanding one bit if we modelled only the damaged ones that could not fly. Once we had built one that could fly, THEN we could go on to model damaged partial function. But before that, partial function is simply underdetermined, admitting of solutions that need have nothing whatever to do with the real thing.

(On the other hand, it would be absurd to expect to capture T3 all at once: It will no doubt have to be converged on by a long series of approximations: toys that eventually scale up to T3, for which the degrees of freedom will be radically narrower than those of the toys along the way. Don’t ask me whether the convergence will be smooth or jumpy: I haven’t the faintest idea; it probably depends on the extent to which the substrate of T3 capacity is modular.)

> And note, I’m not arguing that a richer sensorimotor connection is of
> no significance at all, as a matter of practicality. But I see no
> motivation for human shape or bandwidth when considering symbol
> grounding *in* *principal*, and hence I see no reason to suppose
> that computers necessarily lack grounding.

T3 grounding is needed to distinguish a candidate with a mind from an overinterpreted toy. Turing-Indistinguishability (T3) corresponds to the natural constraints that gave rise to the distinction (between things with and without minds) in the first place. T3 is what guides us in our everyday judgments about other minds, and it is also what guided the Blind Watchmaker in shaping us. I repeat, the "practical" goal is not to feed the right data to a computer; it is to build a system with a mind. A computer executing a program is just an ungrounded symbol system no matter WHAT data it gets or has. For anyone who has understood the symbol grounding problem, it has to be clear that it doesn’t matter what data a computer has, and what program it is running; it cannot be grounded if the connection between its code and what it’s about depends on external interpretability alone. So the real role of grounding is to confer autonomy from external interpretation, NOT to get the right information in there -- for whatever information you get in there is still ungrounded without the autonomous, interpretation-independent causal connection conferred by T3 capacity.

> So, to review, it puzzles me why a human is grounded because of the
> nature of sensorimotor connection to the world, and a robot might well
> be grounded by its sensorimotor connection to the world, and even
> Hawking (or a quadraplegic, or a blind person, or whomever) with
> less-then-T3 sensorimotor connection to the world is still grounded,
> a computer with less-than-T3 sensorimotor connection is just
> emitting squiggles and squoggles.

By now I hope this is answered. Saying more would just amount to repetition.

> The nature of my puzzlement seems to me to be related pretty directly
> to this distinction between "transduction" and "symbol manipulation"
> that Harnad wishes to make. I simply don’t see it. I can’t find
> anywhere adequate definitions that can be used to decide if something
> is one or the other.

trans+duc+er (traenz‘dju:s) n. any device, such as a microphone or electric motor, that converts one form of energy into another. [C20: from Latin transducere to lead across, from trans-+ducere to lead]

com+put+er (km‘pju:t) n. 1. a. a device, usually electronic, that processes data according to a set of instructions. The digital computer stores data in discrete units and performs arithmetical and logical operations at very high speed. The analog computer has no memory and is slower than the digital computer but has a continuous rather than a discrete input. The hybrid computer combines some of the advantages of digital and analog computers. b. (as modifier): computer technology. 2. a person who computes or calculates.

> Consider a robot with a wiring net of 10^10 or so ttl-level bit inputs,
> and another 10^10 or so ttl-level bit outputs. The inputs are
> connected to threshold sensors all over the body for pressure,
> temperature, behind focussed lenses for light level, limb joints for
> proprioception and kinesthesia, and so on. The outputs are connected
> to fibers that react to the ttl signal by either contracting or
> relaxing. The fibers themselves are (of course) attached all over the
> robot’s skeleton, so that turning on more or less of them results in
> the robot’s limbs being moved with more or less force depending on the
> number of fibers in the contracted state.
>
> I think you get the idea. The point is to have something vaguely akin
> to a human nervous system. Harnad’s position seems to be that if we
> consider the subsystem from the sensor-effector terminals *inwards*,
> the voltages on all those terminals are the result of "symbol
> manipulation". But in the analogous subsystem of a human (the CNS
> starting from the next neuron in from direct sensory interaction, and
> on inwards), the firing state of the peripheral neurons is the result
> of "transduction" in the nervous system.

Here are the relevant questions: Does the robot have T3 capacity? If not, chances are it’s just an arbitrary toy. But if it DOES have T3 capacity, WHAT, exactly, has T3 capacity? The system as a whole. If it had no sensorimotor transducers, would it have T3 capacity? No. What would be left if the robot lost all sensorimotor transducers? Who knows? And, whatever that might be, is there any reason to believe that that’s what H-K/S-H? Again, No. Is there any reason to believe that a real brain is a sensorimotor-transducerless core? None whatsoever. And what is a digital computer, no matter what program it is executing? An ungrounded symbol system. This is all based on the homuncular computationalism that ran through Pindor’s original comments (see my Reply to him, which will appear a few postings later.)

> I hope I can predict Harnad’s position well enough to say that we’d
> agree that the subsystems I pointed out (T3 robot with zillions of
> wires from terminal connections "inwards", and human from
> next-neuron-to-sensor "inwards") are *both* ungrounded, and that the
> complete systems (the T3 robot, the human) are both grounded.

I have no idea what kind of a mental partitioning is being imagined here, but I will repeat: Only artificial systems are on trial here. I don't know what's left if you peel off all of a brain's sensorimotor surfaces. But if you remove a T3 robot's sensorimotor transducers and what's left is just a computer, executing some program, then that system is ungrounded (and hence, by my lights, has no mind). The only conclusion I draw from this is that whatever a brain-in-a-vat might be, it won't be THAT.

> If so, then I don't understand why, in humans, simply lowering the
> number of effective neural endpoints by large factors does NOT
> imply ungroundedness, but doing the exact same thing to the robot
> *does* imply ungroundedness.

Because we have no idea what human brains really are, and what they're doing, and how; but we do have a pretty good idea (from symbol grounding and Chinese-Room considerations) what implementation-independent computation and digital computers (and underdetermined toys) are (and aren't).

> Because, after all, what the computer in front of me *has* got is some
> number N of various sensors (keypad sensors on the keyboard, pixel
> sensors on the image scanner, etc), and another number M of effectors
> (screen pixels, speaker voltage levels, etc). (I'll ignore the
> bandwidth of the ethernet connector.) The numbers N and M are
> (of course) ridiculously lower than the analogous numbers for the T3
> robot or the human. But the only difference between the computer and
> the T3 robot I can see is the particular real-world percept or effect
> represented by each of the N+M bits, and the quantity N+M itself.

I don't see a T3 robot anywhere in sight. But once you show me one, I'll show you things it can DO that your computer and its peripherals can't DO. And deprive it of its sensorimotor transduction capacity, and it can no longer do those things.

> I see no recognition in Harnad's model that there is essentially a
> continuum from "a computer" to "a T-3 robot". I see no reasonable way
> to draw a line in this continuum to make a difference in principal
> between the groundedness of a computer, and that of a T-3 robot.

The continuum is not so much from computer to T3 robot as it is from t1 (toy-scale) capacity to T2 to T3 capacity. The only trouble is that symbol systems put the cart before the horse (or try to raise themselves by their bootstraps: take your pick), because the only viable way to scale UP is from the bottom up; and if at any time you eliminate that foundation, the symbols just come crashing down.

Stevan Harnad

The following files are retrievable from directory pub/harnad/Harnad on host princeton.edu

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Harnad, S. (1992) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. SIGART Bulletin 3(4) (October) 9 - 10.

Harnad, S. (1993) Problems, Problems: The Frame Problem as a Symptom of the Symbol Grounding Problem. PSYCOLOQUY 4(34) frame-problem.11.

Harnad, S. (1994) Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. Artificial Intelligence 1: (in press)

Harnad S, (1994) The Convergence Argument in Mind-Modelling: Scaling Up from Toyland to the Total Turing Test. Cognoscenti 1:

----

rickert@mp.cs.niu.edu (Neil Rickert) writes: > >Thinking is indeed the issue. If you were to build a robot with all of >the transducers you think necessary for a T3 system, yet you programmed >it in the way current computer science practice would suggest, I suspect >that the robot would be incapable of thinking, and so would not have a >mind. Quite possibly it would also fail the T3 test. I suspect the >same would be true if the design used the type of neural network software >that has so far been tried.

My model is not a transducer model, but a hybrid analog/connectionist/symbolic model (Harnad 1992, 1993a,b, Harnad et al. 1991, 1994). If it fails to scale up to T3, it fails. If it does scale up to a T3 robot, it succeeds (and it's a grounded symbol system, which is not what current computer science practice aims for). I do not see the substance of your objection.

>Searle was quite right in his intuition about this, even though his >proof does not (in my opinion) succeed. In order to get thinking the >design will need to specifically provide for it. If I am right in this, >then it might also be true that a T2 system which specifically provides >for thinking in its design will also have a mind.

To see where Searle and I agree and differ here, see each of our contributions to Harnad 1993b. To get back into ungrounded symbol systems and T2 would be just to start repeating ourselves, so I'll leave it at that.

>The hypothesis "cognition is computation", at least as I view it, is >really about the nature of what is required to get thinking going. >Questions of implementation independence are not central to the issue. >If I understand you properly, your intuitions are that the right >transducers are the missing ingredient. My intuitions differ from yours >as to what is missing.

You do not understand me properly: whatever can generate T3-capacity is the missing ingredient (and my hybrid model is one candidate, if it can scale up to T3); and you can't have T3 capacity without sensorimotor transduction. A pure symbol system (and implementation-independent implementation of computation, ANY computation) is ungrounded, cannot pass T3, and hence cannot have a mind.

> Nobody doubts that getting from point A to point B on an airplane cannot
> be done when the computation is done in a computer rather than with the
> physical aircraft. However an airplane can be flown to learn its
> aerodynamic properties. And a computer simulation works quite well for

> this. Aircraft designers fly their planes in simulation before they
> build them, precisely to learn about aerodynamic properties.
>
> People who argue that "cognition is computation" are claiming that
> thinking has more in common with solving aerodynamic problems than it
> has with getting from point A to point B. After all, if I think about
> flying to Paris, I still don't finish up in Paris at the end of my
> thoughts.

People who argue that "cognition is computation" confront a lot of problems, among them the symbol grounding problem. And, as far as I know, all systems that we KNOW have minds, also have T3 capacity. And it's PEOPLE who solve aerodynamic problems, USING computers (and interpreting their symbols), not computers themselves. And thinking cannot have something in common with THAT, on pain of infinite regress. (That's the symbol grounding problem.)

>sh >That a digital computer can be reconfigured by its software to be an >sh >implementation-independent implementation of (just about) any and every >sh >symbol system is a physical property of digital computers (and not, say, >sh >planes, or even brains).
>
> Now this is a very interesting claim. I suspect it is probably correct.
> You are claiming that there are some symbol systems which a computer
> can implement, but a brain cannot. You have stumbled into a fatal flaw
> in Searle's Chinese Room. For the AI program which the CR is trying to
> implement might be one of those symbol systems which the brain cannot
> implement, or at least cannot implement in the was the CR attempts to
> do it. If that is the case, then Searle cannot get his room started,
> and his argument fails.

This is no fatal flaw in the CR Argument. Searle's brain is not reconfigured when he executes the T2-passing program, his body is. It is a trivial fact that we are all capable of implementing any computer program that a computer can implement, given enough time.

> Searle's claim is that if the Chinese Room understands, then that
> understanding must be in Searle himself. I disagree with that claim,
> but for the sake of the present discussion, let me assume that it is
> correct. Assume also that my estimated timings are correct. Then,
> based on my timings, Searle should indeed understand. The first
> glimmerings of understanding will occur to him after about 10,000 years
> of continuous operation of the room (because of the slow computation,
> this is equivalent to about 1/10 of a second of normal human time). To
> the best of my knowledge we have no evidence from psychology which would
> deny the possibility of special understandings emerging after 10,000
> years of steady performance of a task requiring meticuluous attention
> to detail. In that case Searle does not have the evidence available to
> back up his claim that he would not understand.

This will have to be my last contribution to this exchange, for the discussion again regressing into sci-fi fantasies about phase transitions into mental space once computation attains the speed of thought. Unfortunately, I continue to find such speculations arbitrary and ad hoc -- but I've already done enough iterations on this in years past.

Stevan Harnad Cognitive Science Laboratory Princeton University 221 Nassau Street Princeton NJ 08544-2093

Ftp host: princeton.edu directory: pub/harnad/Harnad

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Harnad, S. (1993a) Grounding Symbols in the Analog World with Neural Nets. Think 2(1) 12 - 78 (Special issue on "Connectionism versus Symbolism," D.M.W. Powers & P.A. Flach, eds.).

Harnad, S. (1993b) Grounding Symbolic Capacity in Robotic Capacity. In: Steels, L. and R. Brooks (eds.) The "artificial life" route to "artificial intelligence." Building Situated Embodied Agents. New Haven: Lawrence Erlbaum

Harnad, S., Hanson, S.J. & Lubin, J. (1991) Categorical Perception and the Evolution of Supervised Learning in Neural Nets. In: Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology (DW Powers & L Reeker, Eds.) pp. 65-74. Presented at Symposium on Symbol Grounding: Problems and Practice, Stanford University, March 1991; also reprinted as Document D91-09, Deutsches Forschungszentrum fur Kuenstliche Intelligenz GmbH Kaiserslautern FRG.

Harnad, S. Hanson, S.J. & Lubin, J. (1994) Learned Categorical Perception in Neural Nets: Implications for Symbol Grounding. In: V. Honavar & L. Uhr (eds) Symbol Processors and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling: Steps Toward Principled Integration. Acadamic Press.

Date: Mon, 07 Feb 1994 12:19:02 -0600 From: Neil W Rickert

>>Thinking is indeed the issue. If you were to build a robot with all of >>the transducers you think necessary for a T3 system, yet you programmed >>it in the way current computer science practice would suggest, I suspect >>that the robot would be incapable of thinking, and so would not have a >>mind. Quite possibly it would also fail the T3 test. I suspect the >>same would be true if the design used the type of neural network software >>that has so far been tried. > >My model is not a transducer model, but a hybrid >analog/connectionist/symbolic model (Harnad 1992, 1993a,b, Harnad et al. >1991, 1994). If it fails to scale up to T3, it >fails. If it does scale up to a T3 robot, it succeeds (and it's a >grounded symbol system, which is not what current computer science >practice aims for). I do not see the substance of your objection.

You are misinterpreting me when you refer to what I said as an "objection". I was not specifically objecting to anything. Rather, I was suggesting that there are some specific algorithmic aspects of thinking which would have to be in the system design. I am not persuaded that either the symbolic or the connectionist systems that have so far been attempted can meet those algorithmic requirements.

>>Searle was quite right in his intuition about this, even though his >>proof does not (in my opinion) succeed. In order to get thinking the >>design will need to specifically provide for it. If I am right in this, >>then it might also be true that a T2 system which specifically provides >>for thinking in its design will also have a mind. > >To see where Searle and I agree and differ here, see each of our >contributions to Harnad 1993b. To get back into ungrounded symbol >systems and T2 would be just to start repeating ourselves, so I'll leave >it at that.

I am familiar with where Searle disagrees with you. I don't agree with Searle's position on this.

You keep assuming that I want to "get back into ungrounded symbol systems". I have never favored that approach. When I disagree with Searle, I am disagreeing with the validity of his _argument_. You seem to assume that I am professing a belief in the virtues of traditional symbolic AI, when I am not doing any such thing. I agree with Searle's conclusion, but not his argument, when it applies to traditional symbolic AI. I disagree with Searle when he tries to apply his conclusion in great generality to all forms of computation including T3 systems. However, if Searle's argument were a valid proof, then his conclusions would apply to T3 systems. If you think that I have an axe to grind, then I can assure you that the only axe is to argue that Searle's thought experiment does not apply to T3 systems. I have no personal interest in producing T2 systems.

In short, I don't think we disagree nearly as much as you suspect. Our major disagreement over Searle is with whether his argument has the status of a proof. I doubt that we will ever agree on this. It is a question of different background assumptions leading to different intuitions. Searle treats the Systems Reply as a subterfuge designed to evade his argument. For many people in computer science the Systems Reply is completely natural and it is the assumption that Searle himself would understand which we see as the subterfuge. These are different intuitions arising from different experience. These arguments will never be settled by anything less than the construction of an AI system which serves as a counter example to Searle. In my opinion, a successful T3 system would be a sufficient counter example to refute Searle.

>>The hypothesis "cognition is computation", at least as I view it, is >>really about the nature of what is required to get thinking going. >>Questions of implementation independence are not central to the issue. >>If I understand you properly, your intuitions are that the right >>transducers are the missing ingredient. My intuitions differ from yours >>as to what is missing.

>You do not understand me properly: whatever can generate T3-capacity is >the missing ingredient (and my hybrid model is one candidate, if it can >scale up to T3); and you can't have T3 capacity without sensorimotor >transduction.

I understand you completely on this. In other words, the transduction is part of what is missing. What some of us are saying is that there is something else beside transduction that is missing, and it is that "something else" that we are talking about when we say "cognition is computation." But how can this "something else" even be discussed if every attempt to discuss it is always misinterpreted as being an argument against transduction or against symbol grounding.

SUFFICIENT VS. NECESSARY CONDITIONS IN MIND-MODELING

> Date: Mon, 07 Feb 1994 12:19:02 -0600
> From: Neil W Rickert
>
> You are misinterpreting me when you refer to what I said as an "objection".
> I was not specifically objecting to anything. Rather, I was suggesting
> that there are some specific algorithmic aspects of thinking which would
> have to be in the system design. I am not persuaded that either the
> symbolic or the connectionist systems that have so far been attempted
> can meet those algorithmic requirements.

In this context, there is an ambiguity here, and (as will be evident below) the kind of suggestion you are making can be interprteted two ways; one is unimpeachable, but irrelevant, and the other (the one I attributed to you) is, I think, just wrong.

I would be the first to agree that it will surely take a lot to pass T3, and surely vastly more than I've captured in my toy hybrid model. The "more" may be in (1) the implementation-independent computation, in (2) the implementation-dependent transduction, or in some other noncomputational, (3) implementation-dependent structures and processes no one has thought of yet. If all you are saying it that you think the "more" will be in (1), fine, but so what? That is irrelevant to the thesis under discussion here, which is: [C] "Cognition is JUST a form of Computation." That cognition might be PARTLY computation is not at issue -- of course it might. It is irrelevant because it's still true that (1) alone, even with everything missing algorithm provided, is ungrounded, cannot pass T3, and cannot have a mind.

If you agree that C is false, there is no disagreement between us and we are just speculating about possible future details. But even the way you put it here -- "meet those algorithmic requirements" -- looks as if it wants to be construed as follows: Any system that meets those algorithmic requirements will think -- which is just C all over again.

>nr >Searle was quite right in his intuition about this, even though his >nr >proof does not (in my opinion) succeed. In order to get thinking the >nr >design will need to specifically provide for it. If I am right in this, >nr >then it might also be true that a T2 system which specifically provides >nr >for thinking in its design will also have a mind.
> > >
>sh>To see where Searle and I agree and differ here, see each of our >
>sh>contributions to Harnad 1993b. To get back into ungrounded symbol >
>sh>systems and T2 would be just to start repeating ourselves, so I'll leave >
>sh>it at that.
>
> I am familiar with where Searle disagrees with you. I don't agree with
> Searle's position on this.
>
> You keep assuming that I want to "get back into ungrounded symbol
> systems". I have never favored that approach. When I disagree with
> Searle, I am disagreeing with the validity of his _argument_. You seem
> to assume that I am professing a belief in the virtues of traditional
> symbolic AI, when I am not doing any such thing. I agree with Searle's
> conclusion, but not his argument, when it applies to traditional symbolic
> AI. I disagree with Searle when he tries to apply his conclusion in

> great generality to all forms of computation including T3 systems.
> However, if Searle's argument were a valid proof, then his conclusions
> would apply to T3 systems. If you think that I have an axe to grind,
> then I can assure you that the only axe is to argue that Searle's thought
> experiment does not apply to T3 systems. I have no personal interest
> in producing T2 systems.

This again sounds equivocal. If you agree with Searle's conclusion, whether or not you agree with his argument, then how can you hold that "it might also be true that a T2 system which specifically provides for thinking in its design will also have a mind"? If that "provision for thinking" is noncomputational, we have no problem, but then the point is irrelevant. If that "provision for thinking" is just computational, however, then you are contradicting the conclusion you say you accept.

(And, needless to say, "specifically provides for thinking" is itself equivocal, because if whatever "provision" it has indeed makes it think, then it thinks! The question is, what KIND of provision can do that? And the conclusion of Searle's argument, as well as the implication of the symbol grounding problem, is that computation alone is not the right kind of provision.)

Most people, whether or not they agree with its conclusions, fail to understand Searle's argument. I have written enough on that by now so that I am content to leave the rest to Achilles and the Tortoise to work out. But I have to point out that:

(a) It's a mistake to describe Searle's Argument as a "proof." It's a plausibility argument. It remains logically possible that manipulating a bunch of meaningless Chinese Symbols (1) generates conscious understanding of Chinese in the manipulator, or (2) generates another mind, of which the manipulator is unconscious, but that consciously understands Chinese. It is even logically possible that (3) if only Searle could simulate symbols fast enough, some sort of phase transition into mental hyperspace would occur. I've enumerated these possibilities many times, so there's no point talking about Searle's "proof." (On the other hand, clairvoyance, creationism, and the Great Pumpkin remain possibilities too.)

(b) The whole point of Harnad (1989) was that Searle's Argument "does not apply to T3 systems." That was where the T3 terminology came from! So if that's your axe, it's already been swung. Searle's Argument (no proof in any case) is valid ONLY against the thesis that a pure symbol system that could pass T2 would have a mind. Move to T3 and away from pure symbol systems and all bets are off.

> In short, I don't think we disagree nearly as much as you suspect. Our
> major disagreement over Searle is with whether his argument has the
> status of a proof. I doubt that we will ever agree on this. It is a
> question of different background assumptions leading to different
> intuitions. Searle treats the Systems Reply as a subterfuge designed
> to evade his argument. For many people in computer science the Systems
> Reply is completely natural and it is the assumption that Searle himself
> would understand which we see as the subterfuge. These are different
> intuitions arising from different experience. These arguments will
> never be settled by anything less than the construction of an AI system
> which serves as a counter example to Searle. In my opinion, a successful

> T3 system would be a sufficient counter example to refute Searle.

Ah me! A successful T3 system would change the subject, because that is not what Searle's Argument was about. As for T2 and the System Reply: The illusion that the system as a whole has a mind works fine for a T2-passing digital computer for the simple reason that no one can be the wiser (because thinking is safely unobservable). But when Searle himself becomes the WHOLE SYSTEM it becomes a little clearer just what nonsense the belief in C actually commits us to (hence the plausibility argument).

>nr >The hypothesis "cognition is computation", at least as I view it, is >nr >really about the nature of what is required to get thinking going. >nr >Questions of implementation independence are not central to the issue. >nr >If I understand you properly, your intuitions are that the right >nr >transducers are the missing ingredient. My intuitions differ from yours >nr >as to what is missing.
> >
>sh>You do not understand me properly: whatever can generate T3-capacity is >
>sh>the missing ingredient (and my hybrid model is one candidate, if it can >
>sh>scale up to T3); and you can't have T3 capacity without sensorimotor >
>sh>transduction.
>
> I understand you completely on this. In other words, the transduction
> is part of what is missing. What some of us are saying is that there
> is something else beside transduction that is missing, and it is that
> "something else" that we are talking about when we say "cognition is
> computation." But how can this "something else" even be discussed if
> every attempt to discuss it is always misinterpreted as being an argument
> against transduction or against symbol grounding.

What would make it much clearer and less equivocal would be if you made it quite explicit whether you think that "something else" alone could be just implementation-independent computation, so that any and every implementation of it alone would indeed be thinking and be a mind.

To put it even more simply: The issue concerns SUFFICIENT conditions for cognition, not NECESSARY ones. The Computationalism thesis is that computation is sufficient. The symbol grounding problem and Searle's Argument imply that it is not. Coming back that it may be necessary is just a non sequitur.

Stevan Harnad

Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

SUMMA CONTRA GENTILES To would-be refuters or "solvers" of the Symbol Grounding Problem

This is to those who think there is no symbol-grounding problem (or, worse, that the fact that a symbol system is "useable" in the "right" way is somehow a "solution" to the symbol grounding problem):

I think; my thoughts are (1) about something, and (2) systematically interpretable as being about something. 1 and 2 are not the same. A symbol system has 2 but not 1. That's the symbol grounding problem. (If 2 were the same as one, we'd have an infinite regress.)

"Useability," like interprability is just an external criterion. My thoughts are not what they are about simply because you can systematically interpret them as being about what they are about; I differ from a static symbol system like a book or a dynamic one like a computer (running ANY program) in that respect. By the same token, my thoughts are not what they are about simply because you can USE me in some way!

This kind of thinking is just symptomatic of getting lost in the hermeneutic hall of mirrors -- the illusion we fall into when we forget to distinguish what a symbol system REALLY is from what it is that we can interpret or use it as.

Computation is always implemented out of transduction, but the physical details of the transduction are irrelevant to the computation (which could have been implemented in radically different ways -- all still SOME form of transduction). This is the basis of the difference between an implementation-DEpendent "transducer," such as a planetary system or a plane, and an implementation-INdependent computation, such as a virtual planetary system or a virtual plane. Both have "transducers," of course, but the latter two have the WRONG ones (if what one wanted to generate was REAL planetary motion or flight).

With thought, it's rather more complicated, because thought is unobservable (except to its thinker). It is not that thinking IS T3 robotic performance (that would be a silly form of verificationism I would never endorse). My hypothesis is that T3 robotic CAPACITY GROUNDS SYMBOLS, and that thinking is part of the internal activity of a grounded T3 system. The full T3 scale is critical for the hypothesis, because that is what narrows the degrees of freedom from the arbitrary number of ways that you could make a system that could JUST play chess or basketball, to the (presumably much fewer) ways that you could make a system that has our FULL T3 capacity.

So forget about introspections about brains (homuncular digital computers, actually) in vats, and focus on capturing T3. For this you need real sensorimotor transduction (among other noncomputational things: I also think you need a lot of other physical, analog structures and processes that, like sensorimotor transduction, planets and planes, are implementation-DEpendent, but just hold sensorimotor transduction in mind as an example; it will do just fine). A computer without sensorimotor transduction is NOT Helen-Keller/Stephen-Hawking but just another ungrounded symbol system, with neither T3 capacity nor a mind. The time to see how one can scale down a system with a mind to H-K/S-H is AFTER you have found out what it takes to get a system with a mind in the first place! And for that you cannot do without sensorimotor transduction and T3. The only thing I can say a priori about the scaled back version is that it WON'T be just a computer running the right program (because of the symbol grounding problem and Searle's Chinese Room Argument -- which, I repeat, is valid ONLY against a T2-passing symbol system, NOT against a T3-passing robot, because Searle can BE the former by implementing the same symbol system without understanding, whereas he cannot BE the latter, because a robot is not just am implementation-idependent symbol system).

Now that this discussion has grown, and familiar old errors are being resuscitated as if they were solutions, or demonstrations that groundedness is a nonproblem, I unfortunately do not have the time to re-enact the symbol grounding discussion of several years ago. I benefitted from that one, though many thought I was wasting my time patiently trying to explain for the Nth time, from an Nth different angle, what was wrong with a number of hopeful counterarguments to Searle. It was in that discussion (I too was a critic, of sorts, of Searle, yet I knew a bad counterargument when I saw one) that the very term "symbol grounding problem" occurred to me, and, after over a year of exchanges, I had heard it all (many times over), all the would-be lines of objections, and what their flaws were. I say my time was not wasted, because that's what helped give birth to a long series of papers I have since written.

But those papers are now written. It is not that I am referring people to chapter and verse; that is against my nature, and besides, I continue to think that the issues are clear enough and simple enough so they can be re-stated succinctly. So when I first encounter a familiar fallacy or misconception, or symptom of the "hall of mirrors," I reply in situ. It is just the endless rounds of repetition that I alas no longer have the time for (nor, frankly, do I derive any insight from them any more, since new objections do not seem to be appearing, as I had hoped, and the problem -- to me, at least -- is that the objectors have simply not understood [or in some cases even heard] the replies that have already been made, repeatedly).

So with this, I must bow out of this discussion, though I will continue to look in occasionally, just in case, mirabile dictu, a new substantive point is made.

Stevan Harnad

To: John Searle From: Stevan Harnad

John, before I circulate the Chinese Gym piece (which I have only now gotten a chance to read) I want to go on record as urging you to reconsider the argument before launching it.

This piece, in my opinion, would devalue the currency that you have established with the original Chinese Room Argument. The remarkable power of that Argument -- which I have now defended successfully countless times -- is its built-in immunity to any "Systems Reply." Whenever anyone says that YOU may not be understanding Chinese but the "system" might, you point out that you are in fact performing ALL the FUNCTIONS of the system! What FUNCTIONALIST can counter that? They are committed to the irrelevance of the implementation of the function, so they can't say that in the computer version the system thinks and in the Searle version it does not, yet all there IS to the system in the Searle version is you!

This, I think, is what has made the Argument so long-lived and the object of so many obsessive and unsuccessful attempts to refute it. Most of those attempts are various variants of the Systems Reply -- the persisting intuition that even if a part doesn't understand, the whole could. In rebutting this on your behalf, I keep on reminding people that in the Chinese Room there ARE no functioning parts apart from you!

And as to the second most common rejoinder -- also Systems-Reply-inspired -- that the brain has nonthinking parts too, and works as a system -- I keep reminding people that you have nothing against systems, parts, or even systems of which you may be a nonthinking part! Logically speaking, your Argument applies only to "systems" that allegedly understand, but ALL of whose functions you yourself can perform without understanding. Pure symbol manipulating systems of the kind produced by Strong AI are precisely systems of this kind: You yourself can perform all of their functions without understanding, hence they do not understand when they do the same thing.

The coup de grace is when you point out that there is at least one "system" that you DO believe understands, and that's the brain, and it understands because it does have the causal powers necessary and sufficient for understanding, whereas symbol systems do not. You are not even opposed to artificial systems in principle, as long as they have the requisite causal powers, which, again, on the strength of the Chinese Room Argument, symbol systems (and any other system all of whose functions you yourself can perform without yourself understanding) do not!

This network of points and counterpoints is, in my experience, the enduring strength of the Chinese Room Argument against all comers to date. But in the Chinese Gym you give up this built-in immunity to the Systems Reply, leaving yourself expressing the same kind of prima facie skepticism about the distributed "system" of un-understanding bodies there as the god-fearing old lady expresses about both the computer AND the brain!

You don't want to do that! I'm not saying that nets DO have the requisite causal powers to understand; I'm just saying that so far you don't have an Argument that deserves to stand in the same room as the Chinese Room Argument to the effect that they don't. You have only the old lady's skepticism about dumb parts. But that's just what the Systems Reply has been waiting for! If you endorse this as an Argument against connectionism, you'll not only lose there, you'll cast a shadow on the rock-solid ground the original Chinese Room Argument stood on, making it seem that you were making the weaker case -- vulnerable to the Systems Reply -- all along.

There is a way out of this, I believe, and I hope you'll consider using IT as your argument against connectionism instead: It is a fact that despite all the talk about "brain-likeness" and "networks of activation with neuron-like units," etc., all nets, and all their accomplishments to date, have been COMPUTER SIMULATIONS of nets; these, like all computer simulations, are in reality of course just symbol manipulating systems! Yet nets have been proffered as radical alternatives to symbol systems.

Are they? Well, in a sense they are, because they could in principle be implemented as real nets, which are radically different from symbol systems, which can only be implemented as symbol systems (e.g., turing machines or von-Neumann style digital computers, or their functional equivalents, like a Chinese Army with a string of beads). In my view, though, it is not the possibility of this radically different implementation that is the real difference between the symbolic approach and the connectionistic approach. They are actually just a rival set of algorithms within the same basic framework, all of which CAN be not only SIMULATED but also IMPLEMENTED purely symbolically.

Now THAT's the key to resurrecting the Chinese Room: As I argued in "Minds, Machines and Searle," any function that can be completely implemented purely symbolically is vulnerable to the Chinese Room Argument. Unless the connectionists can give a reason why the implementation of a net as a physical net rather than as a computer simulation of a net is ESSENTIAL to its putative mental powers, you -- you alone, not a gym full of boys, which might, just might, have the requisite system properties -- you alone will still be able to perform all the functions of the (simulated) net, just as you were able to perform all the functions of the pure symbol manipulator ("Strong AI"), without understanding!

And I don't think the connectionists will want to argue that there IS this magical essential difference between the two implementations of the same function because, after all, most of them are likewise functionalists, for whom the function is implementation-independent. In any case, if they DO want to say there's an essential difference between a simulated and an implemented net, they will be stuck waving their hands about the function's actually having to HAPPEN in parallel in real time (which you can't ape, even hypothetically) rather than serially, which seems not much better than saying that a function must happen at a certain speed, rather than more slowly. I say hand-waving, because this is not yet a FUNCTIONAL distinction: No functional reason has been given why one implementation of the function should have different causal powers from the other. If the serial, slower implementation delivers exactly the same goods, the burden is on THEM to say why one version should have a mind and the other should not.

Let me close with a plug for the symbol grounding problem: Unlike the hand-waving for the "essential" role of speed or parallelism in implementing a mind, the appeal to transducer function has a much stronger, non-hand-waving rationale: By its nature transducer function

(1) cannot be performed by you, (2) is nonsymbolic, and (3) offers a natural link (via analog representations -- likewise nonsymbolic -- and categorical representations, which CAN be accomplished by a net, but only as a mindless component) to the world of objects to which internal symbols might refer.

Hence transduction strikes at the heart of why pure symbol manipulation may be ungrounded -- and lack intrinsic intentionality, and be vulnerable to the Chinese Room Argument.

So I want to say that the transduction of physical energy from the objects in the world is not only an essentially nonsymbolic function, but it is essential to the implementation of a mind. (The reason why transduction cannot be just an add-on component module, as in the standard Robot Reply, is, as you know, a more complicated story, and more closely linked to my own specific, nonmodular, bottom-up grounding theory, so I won't repeat it here. Suffice it to say that there is a coherent rationale for transduction's essential role in the physical substrate of mind, and hence for its immunity to the Chinese Room Argument, whereas there is no such rationale for parallelism, speed, or even the property of physical continuity.)

And note that nets will need transducers (and analog re-presentations of them) too. We don't know which of the brain's functions are necessary and sufficient for giving it the causal power to implement a mind and which functions are irrelevant to it. The Chinese Room Argument shows that (if the brain does symbol manipulation at all), symbolic function is NOT SUFFICIENT for implementing a mind. The variant I described above suggests that ditto is true for connectionist function: it's NOT SUFFICIENT, for much the same reason. So all I would add is that trandsucer function may be NECESSARY and even primary in implementing a mind, with all the rest grounded in it. Beyond that, at this stage of our understanding, neither I nor you nor anyone else should claim to have given an exhaustive list of the functions that are necessary and sufficient to give a system the causal power to have a mind.

So I respectfully recommend that you jettison the Chinese Gym Argument and instead deal with connectionism by turning the Chinese Room Argument on its head, as follows. Suppose there are three rooms:

(1) In one there is a real Net (implemented as physical units, with real physical links, real excitatory/inhibitory interconnections real parallel distributed processing, real backpropping, etc.) that could pass the Turing Test in Chinese (Chinese symbols in, Chinese symbols out).

(2) In the second there is a computer simulation of (1) that likewise passes the TT in Chinese.

(3) In the third is Searle, performing ALL the functions of (2), likewise passing the Chinese TT (while still not understanding, of course).

Now the connectionists have only two choices:

Either they must claim that all three understand Chinese (in which case they are back up against the old Chinese Room Argument), or the essentialists among them will have to claim that (1) understands but (2) and (3) do not -- but without being able to give any functional reason whatsoever why.

Now the foregoing is an ARGUMENT, in the spirit of the original Chinese Room Argument, whereas the "Chinese Gym" is more like one of those pale Hollywood sequels ("Chinese Room II"), trying to capture the old glory (and its gate), but delivering only the style and not the substance.

Here's a suggestion: If you want a sample of the kind of ironclad defense I could successfully mount against the typical attempted assaults on the Chinese Room -- but a defense that would definitely FAIL in defense of the Chinese Gym, see the next message, which is Dyer's critique of Searle, with my rebuttals.

If despite all this, however, you still want to launch the Chinese Gym rather than the variant I suggested skyward, let me know, and I'll start the count-down...

Stevan Harnad

To: John Searle From: Stevan Harnad

John, here are my comments on your Sci Am paper. In general, I find that it does the job and will get the idea across, but it's rather wordy and redundant, and could use some tightening up. At a few points it's also unnecessarily convoluted. It could also, in general, use a bit more pep, like the original argument. An enthusiasm for the points you're making isn't getting through. Here are the particulars:

> Conclusion 1. Programs are neither constitutive of, nor sufficent for,
> minds.

This an instance of unnecessary convolution and jargon. In plain English this just says a running program doesn't have what it takes to produce a mind. Now there must be some way of saying this that's not quite as vernacular as my way, but doesn't fall into the austere, off-putting and slightly obscurantist jargon of "constitutive of"! [And if you don't do it, the Sci Am editors will, in transforming it to "house" style; but then you risk losing not only the style but the substance of your intended meaning.]

> But from the fact that a system is computational and the fact that it
> is thinking it does not follow that thinking is computational.

This is the first of several places that this is going to cause problems and confusion. Look, there are two notions of "computation" lurking throughout all this business. One is pure symbol manipulation (the syntactic manipulation of meaningless symbol tokens on the basis of formal rules operating exclusively on the (arbitrary) shape of the symbol tokens) and the other is "computation" in the "turing equivalence" sense, according to which EVERYTHING -- every state and process -- "IS" computation. Now there's nothing to disagree about if we mean computation in the latter sense, but nothing much has been accomplished either.

[Yes, I know the two senses are linked: The link is that computation-2 ("turing equivalence") really just means that every state or process is DESCRIBABLE or SIMULABLE by symbol manipulation (computation-1), and that's the ("functional") level at which they're all equivalent. But for this discussion the two senses must be very carefully kept distinct -- as AI people have never bothered to do -- so that it's clear that the distinction has to do with what can be SIMULATED SYMBOLICALLY (answer: just about everything) versus what can be IMPLEMENTED SYMBOLICALLY (answer: lots fewer things than you might expect, e.g., not planes, furnaces, stomachs, transducers, or minds -- for much the same reason in each case: these exceptions are simply not symbolic processes).]

So I urge you to abandon the ambiguous "computational" vocabulary, which will have the reader back and forth between the senses of computation you may mean, and call a spade a spade:

"But from the fact that a system can be simulated by symbol manipulation and the fact that it is thinking it does not follow that thinking is symbol manipulation."

> The answer is: the brain doesn't just \fIimplement\fR a formal pattern
> or program (it does that too), but it actually \fIcauses\fR mental
> events in virtue of very specific neurobiological processes.

Again, as I pointed out in "Minds, Machines and Searle," the word "implements" is unfortunately used in two ways too, for not only can I say that a formal program for symbol manipulation is (1) IMPLEMENTED as a running program on that VAX, but I can also say that a furnace can be SIMULATED with symbols (say, in that program I just mentioned) or it can be (2) IMPLEMENTED as one of those things that gives off real heat.

Here, however, I don't suggest abandoning the ambiguous terminology (because actually "implementation" means exactly the same thing in both cases), but rather pointing out that a computer is the implementation of one very special (and general) type of system: A pure symbol manipulating system. There are other types of systems, however, (like planes, furnaces, stomachs, transducers, minds) that cannot be implemented as pure symbol manipulating systems (although they can be simulated by them). They can only be implemented as whatever kind of nonsymbolic system they happen to be.

I know you like the "causal" part, and I don't disagree with it, but I just don't think it's a perspicuous way of putting it here. It raises unnecessary mysteries (and tempts people to start "counting," as you enjoin them not to do: brain = cause, mind = effect: 1 + 1 = 2).

If instead of talking about what brains do or don't cause you simply talk about the necessary and sufficient conditions for IMPLEMENTING something, then your proposal will sound no more dualistic than it does in the case of airplanes and furnaces: Minds, like airplanes and furnaces, can only be simulated but not implemented by symbol-manipulators. There's no need to say that flying or heating or thinking are "caused" by some other mysterious kind of "stuff" (though that might be true too). You just have to point out that symbol manipulators do not have the properties necessary and sufficient to implement flying or heating or thinking, whereas planes, furnaces and brains do. (And in the case of the brain, it was a reasonable hypothesis, in trying to isolate just which of its functions WERE necessary and sufficient for implementing thinking, that it might have been pure symbol manipulation; it just so happens that, on the strength of the Chinese Room Argument -- and also the symbol grounding problem -- that simply did not turn out to be true!).

So I would say instead:

"The answer is: symbol manipulation is not sufficient to implement a mind; so even if the brain does do symbol manipulation, that can't be the way it manages to implement a mind. Nor would a simulation of all the brain's functions be sufficient to implement a mind; because some of those functions -- the ones necessary and sufficient for implementing a mind -- turn out to be nonsymbolic, just as with the nonsymbolic functions of airplanes, furnaces, stomachs, and transducers that are necessary and sufficient to implement flying, heating, digestion and transduction."

> For our purposes the difference between brains and other sorts of
> computers (remember, since everything is a computer, brains are too)

See, here's the confusion over "computer/computation" again. Use symbol manipulator. And distinguish between the actual symbol manipulation a brain may do and the nonsymbolic functions it may perform that can be simulated but not implemented symbolically.

> But the simulation should not be confused with duplication;

I strongly urge you to abandon the simulation/duplication distinction for the simulation/implementation distinction. After all, the real brain (and the airplane) are not "duplicating" anything. Like the proverbial poem, they're just "being" it. And anything else that manages to think or fly will also BE a mind or a "flyer" (let's not quibble about terminology here). The idea that every implementation of one must have the requisite (necessary/sufficient) properties goes without saying: Is everything that has those properties to be called a "duplication" of everything else that has them? I suggest:

"But a simulation should not be confused with an implementation"

(And, a propos, I hope in Sci. Am.'s "Further Reading" bibliography section you'll consider citing "Minds, Machines and Searle," which is where I spelled out this distinction.)

> all mental phenomena are caused by neurophysiological processes in the
> brain.

Again, do avoid the unnecessary and potentially misleading numerology of counting causes and effects, I would urge you to say:

"all mental phenomena are implemented by [or "in"] neurophysiological processes in the brain."

> Conclusion 2. Any other system capable of causing minds would have to
> have causal powers (at least) equivalent to brains.

Here you could safely say "causal power," as long as you don't insist on the awkward and confusing locution "brains cause minds":

"Conclusion 2. Any other system capable of implementing minds would have to have functions with causal powers (at least) equivalent to those of the functions that make it possible for brains to implement minds."

> but it does imply that any nonbiological system capable of causing
> cognitive states and processes would have to have causal powers
> sufficient to do what brains do.

Here too: "but it does imply that any nonbiological system capable of implementing cognitive states and processes would have to have causal powers sufficient to do what brains do."

> remember: any computation that can be done in parallel can be done on a
> serial machine, and any computation that can be done on a serial
> machine can be done by me in the Chinese room.

Glad to see you decided to use the argument I suggested. Let's watch for a while to see whether the Connectionist community manages to come up with a credible candidate for something you CAN'T do serially that might be essential for implementing a mind. So far I see only this "synchronous/asynchronous" business, but it seems clear that whatever it buys you, you should be able to simulate asynchronously. Stay tuned, and if anything substantive comes up I'll let you know, and you can take it into consideration. For now a lot of people seem to be agreeing that my innocent little question touched a very sore point...

> And of course, as described, there is no way that the Chinese gym
> system could come to attach any mental content to the computational
> processes. Like the original Chinese Room, it takes in Chinese symbols
> as input, and it gives out Chinese symbols as output and we can suppose
> that it passes the Turing Test, but as described [!] there is no Chinese
> thought content in the system. As described there is no way [!] that either
> the system or the people in it could come to learn what any of these
> words means, because there is no way to attach any semantic content to
> the symbols.

Look, as I said in response to this as a candidate for skywriting, you can certainly say all this, but you have to recognize (and, preferably, openly acknowledge) that all these musings about what the system as a whole could or could not do by way of "attaching semantic content" has no more logical or empirical force than the skepticism of the god-fearing little old lady I mentioned before, who thinks there's also "no way," for brains and computers alike.

You have to be careful, even in Scientific American, not to be seen as playing off, or playing into the hands of, Luddites who have a naive skepticism about all things material, and especially computers ("Oh, they just do what they're programmed to do"). You don't want to just say what Grandma already "knows."

Now your original Chinese Room Argument doesn't do that, because it has a bona fide logical (and intuitive) "hook" in the way it has YOU do EVERYTHING the computer does, yet without understanding. The Chinese Gym has no such "hook," and if you think it's of heuristic value to sketch out this kind of many-bodies simulation of what a net does, fine, but make it clear that your skepticism about what there is "no way" for THIS kind of system to do does not have the logical force of your Chinese Room Argument, with its built-in immunity to any "systems reply." (And perhaps you could integrate it better with what follows by presaging that a real argument will come soon, in the spirit of the original Chinese Room Argument.)

> Imagine four rooms. The first contains a computer with parallel
> architecture implementing a PDP program for Chinese language
> undertanding. The second contains a traditional serial computer of the
> sort you can currently buy in a store. It is doing a serial simulation
> of the first and is computing exactly the same function. The third is
> the Chinese Gym, which has both the same architecture and is doing the
> same computations as the first. And finally the fourth is the Original
> Chinese Room which does all the same computations as the first three
> using a serial architecture. Now ask: Which rooms if any understand
> Chinese in virtue of their computations?

I'm glad to see the argument I proposed being used, but the 4-way comparison does becloud the point a bit. A parallel implementation, a serial simulation and a Searle simulation give the logic of the point. The gym, I take it, was just a heuristic device for giving a feeling for what goes on inside a net. I don't think it qualifies as an implementation of true parallelism. Perhaps its partially parallel, but in any case, you've already expressed your doubts (on grandma's grounds) about that one. Why drag it in again here? It's another, partially parallel implementation. So what? What LOGICAL force does it carry over and above the 3-room argument?

By the way, I don't think the first room is best described as a "computer with a parallel architecture for implementing a PDP program." It needn't be a computer at all; and there needn't be any program. The most direct implementation of a net is a net: A system of interconnected nodes that activate and deactivate one another as a causal consequence of the input patterns it receives and the "delta" constraint for modication at each node, etc. The first room implements it, the second room simulates it, and the third room Searle-simulates the second room. All pass the Chinese TT. Now the burden on the neural net community is to make an implementational argument (analogous to flying versus simulated flying) for the fact that although (3) and hence (2) don't understand, yet (1) does. If they don't have an argument, then connectionist nets must accept the same fate as symbol systems: They cannot implement a mind. (Note that I still think there are some unanswered substantive questions here about whether or not parallelism has essential nonsymbolic properties that might be necessary for implementing a mind, as transduction certainly does.)

> Suppose that while I am in the Chinese room, the programmers also give
> me questions in English and a rule book for answering the questions,
> but the English questions and answers are entirely expressed in [a

> code I don't know].

This suggestion by Sci Am's editors is completely trivial and redundant, for the following reason: If you think about it, you'll see that (to a close enough approximation) Chinese itself is ALREADY "ENGLISH," BUT IN A CODE YOU DON'T KNOW: For there is in fact no semantic difference between the proposition (encoded in English) "The cat is on the mat" and the same proposition encoded in Chinese (or in binary English) -- only trivial syntactic differences (about which you are in no position to make a fuss, given that you are aiming at semantic content). To a close enough approximation, both English/Chinese and English/Binary-English are completely intertranslatable; you just don't know the rules (in either case). The thought experiment was not, after all, about whether you could think peculiarly Chinese thoughts (whatever that might be), just whether you could understand the meaning of a bunch of meaningless symbols. The situation is EXACTLY the same for both Chinese and Binary-English: They're just both codes you don't understand. Sci Am may as well give you yet another rule book in Octal-Hungarian...

I'm especially surprised that you chose to explicitly acknowledge the source of this trivial and redundant point in the text and not, for example, my substantive one...

> let us suppose that I decide to interpret the symbols as standing for
> moves in a chess game. And I begin to study with interest which side is
> winning by seeing how the symbols represent different positions in a
> chess game. Now, which semantics is the system giving off now? Is it
> giving off a Chinese semantics or a chess semantics, or both
> simultaneously? And then, suppose there is a third person looking in
> through the window and she decides that the symbol manipulations can
> all be interpreted as stock market predictions. So, she begins to
> interpret the various symbol manipulations as standing for fluctuations
> of stocks that she happens to be interested in. So now, we have one and
> the same set of syntactical manipulations that are giving off three
> inconsistent semantic interpretations. At another window is a fourth
> person with a passionate interest in ballet, and he takes all the
> symbol manipulations to... There is no limit to the number of semantic
> interpretations that can be assigned to the symbols because, to repeat,
> the symbols are purely formal, they have no semantics intrinsically.

Be careful, because you're treading on thin ice here. Although it's true that the systematic semantic interpretation of a symbol system is something that derives from us, it is not NECESSARILY true that a given symbol system is amenable to more than one competing and nontrivially distinct semantic interpretation. The conjecture is rather like assuming that every interpretable system must have multiple (several? many? countless?) consistent "dual" or even "nonstandard" models or interpretations. Well, maybe and maybe not. Or better still, it may depend on the system and the interpretation. In any case, you seem to be implying that you can be sure that the four you mention here would fit the same natural language system: That's surely not true.

You might want to state this more tentatively, or better still, choose a finite example in which it really is true (and demonstrable) that several different systematic interpretations are possible. (But then natural language is not a finite case...)

> Well, part of the answer is that we have inherited a residue of
> behaviorism from the social sciences of the past generation. The Turing
> test enshrines the temptation to think that if something behaves
> exactly as if it had certain mental processes then it must actually
> have those mental processes.

I think there's more to it than this. Writ sufficiently large, the Turing Test, for example, in the variant I propose, namely, the Total Turing Test -- which is not arbitrarily restricted to symbols in and symbols out (which is in any case not in the behaviorist spirit) -- can be seen to boil down to ordinary Empiricism:

After all, what ARE the data available to a psychobiologist? Everything you can objectively measure about the organism. This includes only two things: Structure and function. And function subdivides into internal function (the brain's "behavior") and external function (the body's "behavior"). Now internal function (neuroscience) is very tricky to study, and I don't think its study will lead readily to an understanding of either external function (bodily behavior) OR mental states (which I haven't mentioned yet). For one thing, it is not clear (and may not be clear till doomsday) which internal functions are the ones that are necessary and sufficient to implement either our mental capacities or our behavioral capacities. Moreover, the behavioral ones alone are plenty hard to implement.

So it seems a reasonable strategy for the time being to focus on delivering the behavioral goods only, without worrying about either brain function or mental states. Then, once we've narrowed it down to the functions that are necessary and sufficient to pass the TTT (no mean task), we can worry about fine-tuning them to pass the TTTT (in which the internal behavior of the brain must also be captured). But I think the buck stops there. Because once we've done that, there's nothing left by way of objective empirical data. And we'll have to take it on faith that the wherewithal to pass the TTT or the TTTT has somehow managed to draw in bona fide mental states too. (Besides, we're all TTT "behaviorists" about solving the "other minds" problem every day of our lives.)

> And this is part of the behaviorists' mistaken assumption that in order
> to be scientific, psychology must confine its study to externally
> observable behavior.

Not externally observable behavior, but objective data.

> It is best to see Strong AI as one of the last gasps of this
> antiscientific tradition, for it denies that there is anything
> essentially chemical, biological, about the human mind.

I wouldn't give it that much glory. AI was never a SCIENTIFIC movement; it was just technologists over-extending themselves.

> The mind according to strong AI is completely independent of the brain
> in the sense that it is formal and abstract.

The idea here was not behaviorist either. (God knows, the behaviorists were never formalist; they were antitheoretical, dust-bowl empiricists: "data-ists.") The "symbolic" level was pounced upon because it seemed to have the right properties for capturing the mental level: Believe it or not, the implementation-independence of symbolic functionalism was seen as a way of contending with the mind/body problem, the physical realizability of beliefs, etc.! If anything, the people in AI are naive mentalists, animists, even, who are ready to adorn a toy performance with a fancy mentalistic interpretation at the drop of a CAR or a CDR...

> In this debate what both sides are failing to see is the distinction
> between simulation and duplication.

That's simulation and implementation...

Cheers, Stevan

To: Stevan Harnad From: searle@cogsci.berkeley.edu (John R. Searle)

Stevan,

I have now read your comments on the Sci Am piece. they are immensely helpful. I will certainly give you an extra acknowledgment for the four rooms argument. Sorry it was left out in the draft you got.

In he end I do not think I will use "implmentation" and not "duplication". I see the point your are making, but outside the cogsci ai community "implementation " is still an arcane technical term.

I will send you the final draft before it goes out tomorrow. Someday I will tell you the whole story of efforts to suppress the Chinese Room. But just this one part now.

Originally Sci-Am asked me to do an article for them on the Chin. Room I told them I was sick of the subject but would write about somehting else. They said ok but put in a little bit about the C R. Then when I discovered four other attempts to suppress the CR from getting to the general public, i decided - what the hell I will write the article they originally asked for. So I sent it to them and after sitting on it for a long time, THEY REFUSED TO PUBLISH IT. Well I was plenty annoyed so I protested and got some other peoiple to protest, so they finally changed their mind and said ok, we will publihs it but you have to meet a bunch of objections. Now the redundancy in the version yo rightly complain of is due to my having to meet a lot of dumb objections. And the footnote to them is slightly ironic because i am using one of their arguments against me as an argument in my favor. Any way all this has gone on for years and last spring they said we are going to publish it but only along with a reply buy the Churchlands. I said ok but I have to see the C's reply and have a chance to respond in my article. Three weeks ago The C's article arrived and I was given a deadline to early october to answer them.

Anyway , thats where we are.

And Stevan - many , many thanks. It is always a pleasure to deal with your devious hungarian mind and this is not the first time I have benefited enormously.

Yours in X john

To: John Searle From: Stevan Harnad

John,

Thanks for the kind words. Here are a couple of afterthoughts I had intended to rush to you anyway:

(1) On second thought, instead of "causal powers" (which I think has been repeatedly misinterpreted by many as a kind of animism: "The Force Be With You," and all that sort of thing), I think you will be amazed to see that just the simple term "physical properties [necessary and sufficient for implementing a mind]" will work perfectly well, without misunderstanding or loss of logical force, throughout the text.

For example, with real planes, furnaces, stomachs, etc., you can point out that the critical physical properties that make it possible for them to fly, hear, digest, etc., respectively are clearly NONsymbolic properties. They simply can't be implemented as pure symbol manipulation. Now your Argument has shown that the same is true for understanding: Whatever the critical physical properties are, brains have them and symbol-manipulators don't, on the strength of the Chinese Room Argument. (And calling them physical properties should be especially congenial to you, since you consider mental properties to be a kind of physical property, n'est-ce pas?)

And to block the repeated accusation of brain-mysticism, say explicitly (again) that, whatever these critical properties are, you are NOT claiming that only brains could have them, only that brains DO have them and symbol-crunchers don't.

(After all, an implementation-independent functionalism is possible for NONsymbolic functions too: For example, any furnace must have a physical property of increasing the mean kinetic energy of molecules, and any transducer must have the property of transforming energy from one form into another. That still leaves open a vast class of potential implementations.)

> In the end I do not think I will use "implementation" and not
> "duplication."I see the point your are making, but outside the cogsci
> ai community "implementation " is still an arcane technical term.

Ok. But I must point out that you DO use implementation in the essay already (and that "simulation" is no less arcane), that it's very easily explained, both abstractly and by example, and that it IS what you mean (whereas "duplication" is really ambiguous, since even a simulation is a "duplication" of sorts; and, as I said, all viable implementations of the same kind of thing (flyers, heaters, understanders) are, trivially, "duplications" of one another.)

In fact, in view of the conspiracy to suppress your argument that you describe, you might consider some terminology shifts along these lines precisely to de-fuse prior repeated misconstruals...

> Now the redundancy in the version you rightly complain of is
> due to my having to meet a lot of dumb objections.

And yet, now that you've met the objections (and presumably worn the editors down), this is a good time to clean it up so it'll read well and become a canonical source; as it stands, it's still too disjoint and scatter-shot in its exposition and reasoning.

(If you MUST use the editors' silly binary-english example, you might at least point out that, though cute if you're fond of ironies, it's not a substantive point, because both Chinese and Binary-English are really just encoded translations of English. So it's COMPLETELY redundant with the original Chinese Room Argument. In general, I think minimizing the number of unnecessary sci-fi puzzles and mysteries is desirable in a paper that is basically trying to debunk a simple-minded notion, rather than heap counterfactual fantasy on top of counterfactual fantasy -- as the dungeons-and-dragons generation of AI hackers love to do, and then to get lost in their own dungeons...)

(2) Re: Behaviorism and the TT.

Another point to add, because it clarifies a lot and sets the whole implementation/simulation business in perspective, is the one enormous DISanalogy there is among the functions that you've treated largely equivalently in most of the Argument: (A) flying, heating, digesting, transducing, and understanding, implemented, respectively, as (B) planes, furnaces, stomachs, transducers and minds: The last of these -- minds/undersanding -- is the odd man out in one critical respect that is pertinent to the TT: Note that in each case, to implement a B that displays its respective A (let's call them behavioral capacities) requires physical properties other than symbol manipulation. But in the case of every B except the mind, it is possible to confirm that its respective behavioral capacity has been successfully implemented by objective observation alone. Only in the case of the mind is this not enough: The ONLY way to confirm that you've successfully implemented understanding is to be the mind in question. So that element of uncertainty will always be there in this special case -- whether one uses the TT, the TTT, or the TTTT. (And that's special, and it's called the mind/body problem. It necessarily makes us ALL behaviorists in every single case but one: our own.)

The Devious Hungarian

From: harnad (Stevan Harnad) To: pawlicki@kodak.com Subject: Multiple Personality Disorder

Ted, the short answer is that whereas I believe real people -- who definitely do have real understanding -- can have multiple personality disorder, I am not prepared to believe that anything else can until it is first shown that it has understanding IN THE FIRST PLACE. All this sci-fi fantasy is putting the cart before the horse. And in the case of Searle, I don't believe that memorizing a bunch of meaningless symbols will induce multiple personality; the clinical evidence is that the etiolology involves stronger stuff, like early child abuse...

The trouble is that in freely interpreting the doings of lisp code mentalistically for years, AI/CS people have simply lost their critical bearings. They've forgotten that the FIRST question is whether all this mentalistic projection is any more than just that, projection; only after settling that can one go on to speak about the resulting properties of the mind in question. You're using the peculiarities of these still very suspect candidates (such as the fact that they are stored on disk at night) as REBUTTALS against evidence that they don't have minds! That kind of reasoning just isn't valid. It's called presupposing the consequent. Panpsychism is the wackiest and most extreme version of it, but hackers do it with code all the time...

I agree that history is irrelevant, and if and when a synthetic mind (one that really understands, as Searle does) is ever successfully implemented, it can come into existence as a mature adult on Tuesday; but it will have to account for Monday anyway, which will consist either of illusory memories, a black-out, or a frank knowledge of the fact that it is a synthetic (though genuine) mind, created Tuesday.

This, however, could all be SIMULATED in a purely symbolic NON-mind too (which, I repeat, on the strength of Searle's Argument, would NOT understand, even though it supports our mentalistic projections); I brought it up only to play the sci-fi game for a while. None of these counterfactual sci-fi musings are arguments, or decisive one way or the other. They just milk intuitions and counterintuitions. As I wrote elsewhere, I don't believe that a pure symbol cruncher could successfully pass the TT in the first place, because of the symbol grounding problem.

Stevan Harnad

To: Stevan Harnad From: miken@ai.mit.edu (Michael N. Nitabach)

Subject: Intensionality, Semantics, and Symbol Grounding

It appears to me that the essence of the symbol grounding problem is: "What is the mechanism in virtue of which the contents of a mind refer to events in the external world?" This issue can be divided into two sub-issues. First, is the problem of intensionality: "How do the primitive symbolic elements of my mental representations come to refer to entities external to the mind?" Second, is the problem of semantics: "How do propositions built of these primitive symbolic elements come to be true or false?" I think that an answer to the question of intensionality, coupled with an understanding of the nature of the computations on symbols performed by the mind, will provide an answer to the problem of semantics. That, however, is not the issue I wish to discuss here.

The purpose of this letter is to solicit from you a more detailed description of your theory of "symbol grounding based in sensory transduction." (If you have addressed this issue in detail in previous letters, could you resend them to me?) First, I need to understand the specific meaning of "transduction" as you use it. I see this term as referring to the process of transforming patterns of environmental energy into symbols. Based on this definition, it seems to me, e.g. from the following quote, that you are treading on a narrow conceptual ridge between your acceptance that the mind is a symbol processing device (albeit with a very special grounding for at least some of those symbols), and that its symbols are grounded through sensory transduction.

>Its primitive >symbols are grounded by their resemblance and causal connection to the >real physical objects they pick out and stand for. SH]

I don't understand in what sense you can mean that a symbol "resembles" a physical object. To my mind, the very essence of a "symbol" is that it "refers" to an object, but does not resemble it. I don't think that a symbol, taken in your way, can do the double duty of, on the one hand, having the appropriate syntactic properties so as to take part in mental computation, and on the other hand, "resembling" a physical object. The relationship between a symbol and its referent is wholly arbitrary; that is why we have the symbol grounding problem. I see an affinity between your view of symbol grounding, and that of a verificationist semantics coupled with sense data. Verifica- tionism is to be taken as the notion that a symbol refers to the extent that we possess a procedure for verifying that its presumed referent is presenting itself to our senses. Your idea that symbols resemble their referents seems identical to that of the classical sense data theorists. Could you expand on the relationship of your theory to these theories?

My point here is not to present my own answer to the symbol grounding problem; I don't think we are near to possessing such an answer. I do agree that it will lie in some very complicated causal interactions between entities and their environments, involving motor as well as sensory processes. I don't think, though, that a simple appeal to sensory transduction will work. For example, your notion depends on the idea that symbols which are not "primitive", i.e. which are not themselves directly grounded in sensory transduction, must be constructed out of such primitive symbols. The actual state of events appears to be that 1) the number of symbols which are not primitive in your sense is very large relative to the number which are primitive, and 2) success in reducing these complex symbols to a primitive basis has not occurred, and not through lack of trying (witness the failure of verificationism in semantics, behaviorism in psychology, and logical positivism in philosophy of science).

With regard to Searle's view of the origin of intensionality, I can only comment based on a single one of his texts: "Minds, Brains, and Science." It appears to me that, except for his "special pleading" in the case of homo sapiens, his argument implies that *no* purely physical entity can possess intensionality. I know that this is a popularised version of his theories; I would appreciate it if you could lead me to some of his more technical work.

Yours truly, Mike Nitabach, MIT, Brain & Cognitive Sciences

From: Stevan Harnad To: ?

Thank you for your thoughtful message about the computational, algorithmic and implementational levels of function.

It is clear that these three "levels" can be distinguished and are useful to the theorist, but one can still ask whether they are "levels" in the same sense and spirit as the hardware/software distinction (and the hierarchy of compiled languages above it).

In reading Marr, I have always wondered why he chose to call the "computational" level computational, since it involved no computational specification yet. It would have been much more perspicuous (and may have avoided the kind of ambiguity that I think is lurking here) if he had called it instead the level of the "performance capacity" or the "performance regularity." (By "performance" I don't mean the odd man out in the Chomskian "competence/performance" distinction; I just mean what the system is ABLE TO DO.)

Now even the distinction between performance capacity and performance regularity is a bit problematic, because, to use your examples, I could describe "being able to respond to depth from texture" and "being able to conjugate English past tense" as the performance capacity, whereas "being able to respond to objects as more distant if the input texture grain decreases" and "being able to add '-ed' to regular verbs and rote forms to the irregulars" would be "performance regularities." Then what about the capacity to detect texture? or to detect regular verbs?

It seems clear that what's really going on here is a more and more detailed specification of the performance capacity itself: It's the answer to the question "what can the system do?" at greater and greater levels of specificity. One is inclined to think that the answer to the other question -- "how does the system do it?" would be completely independent, but it clearly can't be. "How" is also amenable to a hierarchy of specificity, all the way down to a complete specification of the structure and function of the hardware (including the software it's running, if it's a programmable computer). And some of the more detailed "what" answers are already answers to the less detailed "how" questions.

Now, all of this having been said, is there still room to carve out three nonarbitrary levels here? I think not. I think that the "algorithmic" level is just a hybrid, fixing a certain performance description (a "what") and then applying applying an implementable regularity (a "how") that happens to be specific enough to be implementable in some hardware so as to generate the performance described.

We all know the elementary operations a turing machine actually performs: It just manipulates symbols. That's computation -- and therefore that's the true "computational level." Now an algorithm is just a higher-order symbol string, which is eventually implemented in terms of the elementary turing operations. There may be more than one algorithm that generates the same I/O performance, and there may be more than one way to implement the same higher-order algorithm computationally, and there is certainly more than one way to implement the computations in hardware. And these distinctions probably do matter in modeling performance and especially brain function. But from the functional point of view, there are only three levels here:

(1) What the organism (system) can do, in an I/O sense (the performance level).

(2) The algorithm(s) that can do what the system can do, given the same I/O (the formal or functional level).

(3) How the algorithm is implemented (the implementational level).

(2) and (3) correspond to the usual functionalist's software/hardware distinction, including the usual levels of higher-and lower level programming languages and architectures. (1) is just performance description at various scales of specificty.

Stevan Harnad

To: Stevan Harnad From: miken@ai.mit.edu (Michael N. Nitabach)

Subject: Re: Parallelism: Simulated and Real

In your letter of Oct. 4, you state:

>I know what the computational (software) and the implementational >(hardware) levels are, but I'm not sure what an algorithmic "level" >would be. (It seems to me that for Marr it was just a useful >conceptual or methodological level for the theorist in thinking >about a problem in AI.)

I think that there is a useful way to look at the three levels of analysis that Marr proposed that makes clear what the algorithmic level might be. I think that the computational level is definitely not the software level, but in a sense the software functional specification level. The algorithmic level would be what you call the software level -- that is, the software is the specific algorithm that is used to realize a particular computation. I agree that the implementation level involves mainly considerations of hardware.

An example will make this clearer: Consider the problem of deriving three dimensional depth representations of a view from the texture gradient of the two dimensional image. The computational rule is that, all else being equal, distance of an object increases as the grain of its texture decreases. This *is* the analysis of the system at the computational level; it answers the question, What is the rule (mathematical, linguistic, logical, etc.) being followed in the performance of a certain task? The algorithmic level involves the analysis of things such as, How do we calculate the gradient of the texture grain across the image? Finally, the implementation level involves such details as, Do we use floating point, or integer, values to represent the quantities over which we calculate.

Another example, this time involving language: The grammars that linguists create *are* computational theories. So, the rule for conjugating the past tense is: Add '-ed', unless the verb is exceptional, in which case, use the memorized past tense form. This is a computational theory of the past tense formation in English. Only the algorithmic analysis allows for the consideration of questions such as, How does one distinguish a regular from an irregular verb?

(Note: The above two examples are gross simplifications of the actual computational theories mentioned. This was intentional; what is important for my present purpose is the form such theories take, and the entities they deal with, not their actual content.)

This way of analyzing levels of cognitive information processing has consequences for the empirical study of cognition. Observation of a system under its normal operating conditions will yield information about the computational level of information processing, but it will not provide insight into the algorithmic or implementation levels. This is because the algorithms and implementations used tend to take advantage of the particular normal operating environment of the organism so as to simplify and make more efficient their realization of the computational rule. Almost by hypothesis, these simplifications will *not* be manifest as deviations from the computational rule, if the system is observed under conditions where these lower level assumptions are satisfied. Only when the system is placed in situations where the algorithmic assumptions are violated, can we make inferences about the algorithms (and implementations) used, based on the particular nature of the system's failures.

That is, as long as the system's algorithmic assumptions are satisfied, the system acts as though it is a perfect realization of the computational rule, and we have no information about the particular algorithm used. When those algorithmic assumptions are violated, the system makes errors of one sort or another, and the exact form of those errors can yield insight into the algorithms and implementations used. That is why visual illusions have been so useful to psychologists trying to understand the algorithms used in visual information processing. This also explains why presentation of run-of-the-mill, easily understood sentences, while giving insight into the rules of grammar (to the extent that they are easily understood and accepted as grammatically correct) will not give insight into the parsing algorithms used in comprehension. Only when we push the system with difficult or agrammatic examples do we gain insight into the algorithmic level, through the character of the performance under these unusual levels.

(Please feel free to redistribute this letter to anyone that may be interested.)

Michael Nitabach, MIT, Brain & Cognitive Sciences

From: Stevan Harnad To: Ted Pawlicki

You wrote:

> My point is that whatever the program is (be it a word processor or a
> NLU system) it is still that thing when rolled out to archival tape at
> night.

And Searle’s point is that symbol manipulation alone cannot implement a mind, whether it is stored inertly on a tape OR running actively on the symbol cruncher. One can’t reply to his Argument with the PREMISE that it can.

> you have some problem with thinking that a process can have an inactive
> state which can be stored on some inert media

Actually, I don’t. I would even be prepared to believe that the activation pattern on the surface of a transducer (a necessarily NONsymbolic device) could be stored digitally and then restored, via transducers, to the transducer surface. However, I don’t consider any of the stored or symbolic states to which the data can be reversibly transformed to amount to transduction. And all of this is quite consistent with Searle’s demonstration that symbol manipulation alone cannot implement understanding.

Look, the gist of this is simple: What you store on tape is symbols. The symbols can be transformed into something nonsymbolic. Searle’s Argument is directed only against the symbols, and any implementation of symbol manipulation, and symbol manipulation alone, that is claimed to have a mind. A hypothetical pure symbol manipulator that passes the Turing Test is such a one. Try not to commute freely between symbol manipulation and other processes (because the nonsymbolic processes are immune to Searle’s Argument), and confront squarely the case against pure symbol manipulation.

> the "multiple personality extension of the system reply" claims that
> some one else is home, but Searle is simply unaware of it (him/her).
> And, as I wrote earlier, given our current understanding of
> multiprocessing operating systems and our current observation of
> multiple personality disorders, such an extension to the system reply
> is reasonable.

We’re beginning to repeat ourselves here. I followed your logic the first time, and my reply was that you’re not entitled to help yourself to MULTIPLE personality phenomena until you’ve shown you can deliver a PRIMARY personality. We know brains can do that, but whether symbol systems can or cannot do that is what’s at issue here. So don’t presuppose the outcome. And no matter how much "personality" one can project onto a multiprocessing system in one’s informal moments, that's just a bunch of gratuitous mentalistic overinterpretation until an argument or demonstration otherwise is forthcoming. The only Argument I see so far is Searle's. The rest is just sci-fi and naive analogy. Maybe besides having multiple personality, multiprocessing systems also have anxiety disorders, panic attacks and sudden waves of depression...

> This does not require any mentalisitic anthropomorphism of code. All it
> requires is the acceptance of the TT axiom of the Chinese room
> argument. If this axiom is not accepted (at least for the purpose of
> discussion), all debate on the Chinease room argument is moot, and any
> defense of it is pointless.

Interesting conception: Demonstration by axiom! But you see, Searle was setting out to show that even if symbol crunching alone COULD pass the TT (as I doubt), then it would not understand (because of his Argument). You reply that we must accept that it would understand as an axiom. Well then discussion is indeed pointless...

> up till now the only argument I have heard against the multiple
> personality extension of the system reply is that "it could not pass
> the TT". I would be very interested in hearing of any others.

You've heard others. Here they are again: You can't help yourself to the known multiple-personality potential of known minds to buttress the case for something that has not only not yet been shown to be a mind, but has just been shown NOT to be (by Searle). And if that's not enough, let me repeat that I simply do not believe that memorizing a bunch of meaningless symbols could induce multiple personality in anyone, including Searle, so I don't believe there's another "system" lurking in him. And I'm sure that anyone who hasn't spent too long overinterpreting what you can do with symbol code would readily agree.

Stevan Harnad

From: harnad (Stevan Harnad) To: pawlicki@Kodak.COM Subject: The Logic of Multiple Personality

To: Ted Pavlicki

You wrote:

> The discussion on multiprocessing systems and multiple personalities is not
> intended to support the claim that the system understands. It is intended to
> demonstrate that independent processes can coexist on a single device without
> sharing data.

The existence of neither multiple processing nor multiple personality was ever in doubt. It's multiple personality in SEARLE after memorizing a lot of code that's in doubt.

> The only point made by the multiprocessing discussion is that
> Searle's mind (his natural one) may not be able to understand all the
> processes running in his body. (It does not make any statements about

> the properties of these systems, or what types of experience could
> induce them, it only points out that they can exist. (No mental
> projection or overinterpretation is involved here). If such processes
> can exist, then the argument based on Searle’s introspection is
> invalid.

That Searle COULD have another mind follows purely from the existence of multiple personality syndrome. Multiprocessing has nothing whatsoever to do with it; neither does the irrelevant fact that we don’t understand how our brains work. To think multiple processing has something to do with it IS to overinterpret multiprocessing mentalistically.

> you are using Searle’s conclusion as evidence in support of his argument

No, I was quite careful; I used it only as evidence that symbolic AI cannot help itself to multipersonality phenomena. The logical order is this:

(1) Symbolic AI claims symbol system X understands because it passes the Turing Test. (For the time being, we can’t be sure whether this is true or false, because we’re not symbol system X, that computer over there is, so maybe it understands.)

(2) Searle does everything system X does; he becomes an implementation of system X, just like that computer was, and lo and behold, he doesn’t understand. (Now we do have evidence that X doesn’t understand.)

(3) Now someone tries to say, "But wait, there’s such a thing as multiple personality, therefore there may be another mind in there that Searle is not aware of that does understand."

(4) My point: So far, the only kind of system that has been established to be able to have MULTIPLE minds is the kind of system that we already know ito be able to have ONE mind (at least) to begin with: the human brain. The only evidence we have about symbol systems (Searle evidence) is that they DON’T have a mind. You cannot counter this evidence that they don’t have a mind by citing evidence that systems that do have a mind can have more than one!

It’s as if I give you evidence that not-P (2) and you counter that, because P implies Q (3) and Q implies not-not-P, therefore P!

>> let me repeat that I simply do not believe that memorizing a bunch >> of meaningless symbols could induce multiple personality in anyone, >> including Searle, so I don’t believe there’s another "system" lurking >> in him. And I’m sure that anyone who hasn’t spent too long >> overinterpreting what you can do with symbol code would readily agree. >
> I understand the statement of what you believe. I am unable to
> disprove it. On the other hand, neither have I been presented with any
> reasonable proof or evidence for it. A lot of people believe a lot of
> things. What I am trying to get at is the reason you believe what you
> do.

I hate to say it, but this has all the exquisite logic of special creation! What, after all, is the null hypothesis here, that everything has a mind until "proven" otherwise? Its counterpart is that creation (or any other religious or metaphysical cosmology -- "the universe is a tap-dancing camel" -- is valid until "proven otherwise).

Well, my null hypothesis is that only people have minds until there is STRONG evidence otherwise, that only systems with at least one mind can have more than one mind, and that all evidence suggests that memorizing symbols does not figure in the multiple etiology of multiple personality disorder. The evidential burden is not on me to "prove" that memorizing symbols could not induce multiple personality disorder any more than it is to prove that running circles around a gymnasium couldn't do so. But I'm prepared to keep my eye out for reports in the Journal of Personality Disorders...

Stevan Harnad

From harnad Tue Oct 10 11:31:43 1989 To: pawlicki@Kodak.COM Subject: Sci-Fi and the Counterfactual Stratosphere

Ted, I think we're really repeating ourselves now. I don't think it's possible for a person to induce multiple personality just by memorizing and manipulating meaningless symbols; you do. Hence I don't think that that possibility counters Searle's evidence against understanding in a system that passes the TT purely by manipulating meaningless symbols; you do. There's already one layer of counterfactual conjecture involved in supposing that the TT could be successfully passed till doomsday by pure symbol manipulation in the first place; if you want me to add that IF this were possible, AND IF this induced multiple personality THEN... Fine, but then we're so high up in the counterfactual stratosphere that only sci-fi questions are being settled.

There we have it. What more can we say?

Cheers, Stevan

ON SEARLE'S "INTRINSIC INTENTIONALITY" AND THE SYMBOL GROUNDING PROBLEM

From: Stevan Harnad

Michael G Dyer wrote:
> 1. Searle's "grounding"
>
> Having read Searle and having recently heard a talk he gave at UCLA, I
> do not think Searle thinks symbolic systems lack intentionality because
> they are not grounded in perception. His Chinese box arguments are
> confusions of levels (of where the intelligence resides) and the
> inability to accept that intelligence may arise from the component
> interactions, independent of the types of components used.

You are apparently satisfied with the "Systems Reply." I definitely am not. Searle has repeatedly said that it's not that he doesn't believe intelligence arises from "component interactions," because he does believe that, for example, it arises from the component interactions in the brain. Nor is Searle committed to any particular kind of component type. He just holds (correctly, in my view) that his argument has shown that certain types of system (pure symbol systems, as it turns out, in my analysis) are the wrong types of system to exhibit intrinsic intentionality; their intentionality is all parasitic on our interpretations.

[In my own crtitique of Searle -- Harnad, S. (1989) "Minds, Machines and Searle" Journal of Experimental and Theoretical Artificial Intelligence 1: 5 - 25 -- I try to sort out systematically what Searle is right and wrong about and why, and it all boils down to the Symbol Grounding Problem.]

> In his Chinese box argument he forces the human to be a piece in an
> intelligent system and then asks that human to introspect, or he asks
> the human to act as an interpreter of an intelligence system and then
> again introspect. Instead, he should have been asking the entire "box"
> to introspect. Only then is he going to get an interesting answer.

This misses the essence of Searle’s Argument: The person is not just being a piece of a system, HE’S DOING EVERYTHING IN THE SYSTEM. Whatever function you want to point to that a symbol-manipulator is performing in order to pass the Turing Test, Searle can simply do it himself. Introspection is the right test for whether the mental state that is being attributed to the system he is simulating is really there under conditions where he is performing every last function the system is performing. That’s why Searle says he doesn’t see much left to hang a mind on in the walls and ceiling of the Chinese Room and the chalk on the blackboard. You’re tilting at windmills if you want to insist that THAT system has a mind. It’s not that Searle is claiming NO system can have a mind, even one of which he is himself a component. He’s just claiming that THAT system -- the pure symbol-cruncher under discussion in the Chinese Room Argument -- has no mind.

And he’s simply right about that, according to me. But, as I show in my paper, it turns out that a very simple variant is immune to Searle’s objection: One in which the candidate system’s functions -- the ones all of which Searle is committed to performing himself, if his argument is to go through -- include certain nonsymbolic functions, in particular, sensory transduction. If, for example, the physical system that implements the mental state of seeing something -- that state we all know whether or not we’re in just as directly as we all know whether or not we’re understanding Chinese -- includes transducer function, then Searle cannot simulate it Chinese-Room style. For either he will have to get only the transducer’s output (in which case it’s no wonder he’s not seeing, since he is not performing all the functions the system he is simulating is supposed to perform) or he will have to BE the transducer, in which case he will be seeing after all.

The reason Searle’s argument works with understanding and fails with seeing is that the candidate system for understanding was a purely symbolic, whereas the candidate system for sensing was not. It was this immunity of a seemingly trivial function (transduction) to Searle’s Argument that confirmed my hunch that sensory grounding may be critical for modeling the mind. (And of course I believe that the place to ground understanding is in sensory function.)

> To my knowledge, Searle has never said the following but I will guess
> that if he were asked whether or not something has to be "alive" to
> have intentionality, I suspect that Searle would say that it does
> have to be (I also suspect that Searle would vehemently deny that any
> future system now being built under the term "artificial life" will
> ever really be "alive" in any sense Artificial Life>)

And I think he would be agnostic about whether or not something had to be "alive" (by which I assume you mean a natural biological creature that is not dead) to have a mind. For example, I doubt Searle would have any problem with artificial neurons, as long as they had the same causal powers as natural neurons. But Searle's extracurricular beliefs are irrelevant. Insofar as they are based on the Chinese Room Argument, he has given strong reasons for doubting that mere symbol-crunchers ("Strong AI") can have minds. The rest is moot.

> I think the "grounding" Searle wants really has to do with the units of
> a system being 'grounded' in living tissues (i.e. real neurons, all the
> way down to quarks and other beasties postulated of reality).
> In contrast, most AI researchers think that information processing in
> humans is only loosely coupled to what neural tissues are doing (i.e.
> the processes keeping the cell alive are separable from those being
> used to process information of a cognitive nature).

I repeat: although Searle hasn't said so explicitly, I will: The only kind of artificial system that is vulnerable to his Chinese Room Argument is a pure symbol manipulating system. When it comes to nonsymbolic functions, all bets are off. Searle mentioned in passing that a system with a mind must have all the causal powers of a real brain, but I'm sure he would agree that many of the causal powers of the brain may be irrelevant to its having a mind (as Stephen Hawking's tragic case shows so dramatically), and that all he can really insist on are those causal powers of the brain that are necessary and sufficient for having a mind (which begins to sound a little tautological), and that any system, natural or artificial, that had those, would have a mind. Apart from that, Searle simply points out, correctly, that the kind of system he can simulate in the Chinese room (a pure symbol system) clearly does not have a mind, whereas the brain clearly does.

So the moral is simply that in the attempt isolate that subset of the brain's function (or any other possible mind-supporting system's function) that is necessary and sufficient for having a mind, a pure symbol crunching module turns out not to the right subset!

> I've never heard Searle talk about the importance of grounding symbols
> in perception. If I'm wrong, please point me to his relevant paper(s).

Neither have I. But *I've* talked (and written) about it. And whereas I first thought I was disagreeing with Searle, he assured me that we were largely in agreement. So there you have it.

> If we draw an analogy (dangerous, but permit me) to architecture, my
> guess is that most AI types would say that a building is e.g. Gothic
> as long as it's in the Gothic style, even if each brick is made of
> plastic, while Searle would demand that the actual bricks be made of
> the right clay and maybe even in the right century.
> Clearly both the plastic brick and clay brick buildings have different
> behaviors under stress, but at another level they are both Gothic. It
> depends on whether you think intelligence has to be 'grounded' in the
> clay or if it is the way the elements are put together that counts.

The problem is that "Gothic Style," if it's anything at all, is an objective, observable property. (So is behavior under stress.) In principle we can all agree that any structure that has certain geometric properties, irrespective of any other properties, is Gothic. Unfortunately, this is not true of understanding (or seeing, or any of the other subjective properties of having a mind). For even if a system has all the objective, observable properties of a system that has a mind, it may not have a mind. And that can be true as a simple matter of fact, not just a metaphysical possibility, as Searle's Chinese Room Argument shows.

But here I actually part ways with Searle. Searle rejects all versions of the Turing Test, very much for the above reasons, whereas I reject only the traditional symbolic version of the Turing Test (symbols in, symbols out, and presumably just symbol-crunching in between), because of the Symbol Grounding Problem. The robotic version -- what I've called the Total Turing Test (TTT) -- is good enough for me, especially since that's all I can ever have to go on even with my fellow human beings. Appearances could be deceiving there too (as the "Other Minds Problem" reminds us), but as far as I'm concerned, Total Turing power is power enough to convince me.

Besides, no one knows what the relevant causal powers of the brain are anyway, so that certainly can't be the way to settle the matter. The brain's equivalent of your clay's "behavior under stress" may or may not be one of the properties any system must have in order to have a mind. To put it another way, it's not clear how much of the TTT will have to include not only the body's behavior, but the brain's. Both, after all, are objective, observable properties. The more you capture, the better your grounds for confidence. I just think (total) bodily performance capacity is probably a sufficient constraint.

> The issue is not whether or not future AI systems will have HUMAN
> intentionality, but whether they will have Searle's (elusive)
> intentionality that's good enough, say, to quarantee them human
> rights.

Searle's "elusive" intentionality is the only kind there is. It's the stuff the systems that really see or feel or understand [English or Chinese] have (and know at first hand that they have, because they know exactly what it feels like to see or feel or understand [English or Chinese]) and that systems that only act exactly as if they see or feel or understand [English or Chinese] don't have (but don't know they don't have, because there's nobody home). It's just that I don't happen to believe you can get a system to ACT (TTT) exactly as if it sees... etc., without its really having a mind.

Pure symbol-crunchers, on the other hand, cannot by their very nature have ALL the requisite powers, because some of the powers are nonsymbolic -- and, according to my theory, the symbolic ones must be (nonmodularly) grounded in the nonsymbolic ones; so an autonomous symbol crunching mudule will have no mental powers at all.

> 2. Harnad "Grounding"
>
> But Searle's 'grounding' is different (I assume) from the way you use
> the term "grounding". You have pointed out (correctly, I believe) that
> the relationship of symbols to perceptions is a more important and
> fundamental problem than people may have realized.

And I tried to show above exactly what the connection between my grounding problem and Searle's intrinsic intentionality problem is.

> AI researchers have argued that one can go a long way toward
> intelligence without worrying about grounding (e.g. we imagine that we
> are working with machines that are like Helen Keller). The problem is
> that even Helen Keller had the (incredibly complex) sense of touch and
> movement through the environment.

Exactly. I know of no one -- certainly not Helen Keller or Stephen Hawking -- who is even remotely like an autonomous symbolic module. There is no such nonhuman species either. The symbol grounding problem may be the reason why. And of course the reason AI could go as far as it did with its mind-like feats of symbol crunching was that the symbol-crunching is always parasitic on our human interpretations: From a TTT standpoint, AI's feats are all mere toys.

> As we try to figure out where symbols come from, we will discover that
> the grounding problem is major and will effect the nature of the
> symbols we build and the way we manipulate them.

Especially if we have TTT ambitions.

> 3. Connectionism and Systematicity
>
> Distributed representations aren't all the way to systematicity yet,
> but there are a number of researchers that have been building methods
> for dynamically forming distributed representations that act a lot like
> symbols, yet still have a "microsemantics". For instance, see
> Miikkulainen and Dyer in Intern Joint Conf on Neural Networks (IJCNN),
> Wash D. C. 1989. and in Touretzky, Hinton, Sejnowski (eds.) Proc of
> Connectionist Summerschool See also Pollack in Cognitive Science
> conference 1988 & 89, See Dolan' s UCLA dissertaion on Tensor
> manipulation networks. See also Lee, Flowers, Dyer in 1989 Cogn Sci
> conf., See Sumida and Dyer in IJCAI-89.
>
> Of course, once we get started moving these distributed representations
> around like symbols, then Fodor et al. can use the argument that it's
> "just" an implementation. I think we do need symbols and, unlike
> Chuchland and Sejnowski, I don't expect high-level cognition without
> them. But automatically forming symbol representations through the
> symbolic and perceptual tasks being demanded of them will give us more
> robust intelligent systems and explain currently unexplained
> phenomena.

I too think that symbols enter into cognition at some level. I just think they have to be grounded bottom-up in nonsymbolic function. And the difference between a pure symbol crunching system and a nonsymbolic system that can climb to a symbolic level by learning through sensorimotor interactions in the world is just the difference between theft and honest toil. I also think that the requisite nonsymbolic function will include a lot of transduction and pure analog processing (not just connectionist networks).

And even when a symbolic level is "trained up" in such a hybrid system we will still not have a pure symbol system, because pure symbol manipulation is based only on the arbitrary shape of the symbol tokens, whereas in this "dedicated" symbol system the symbol tokens would also be grounded (via feature-detecting neural nets) in the sensory projections of the objects to which they refer. I predict that there will be no isolable, autonomous symbolic module in a grounded hybrid system with TTT power.

> One problem with symbolic implementations is that they use something
> like ASCII -- i.e. arbitrary, static and predetermined -- forcing us
> into the use of pointers where maybe they're not needed. What we want is
> something that's formed from perception (plus lots of innate wiring
> that extracts features from those perceptions -- that wiring we get
> "for free" from evolution) and modified by interacting with other
> "distributed symbols representations" (DSRs). Instead of ASCII for CAT
> we get a DSR that reconstructus images of cats moving, plus
> reconstructing abstract information about cats when we need that type
> of information. Take a look at Miikkkulainen & Dyer in IJCNN-89 for
> how this can be done. (Notice that although the perceptual part in
> this system is incredibly primitive it's actually being used to help
> keep bindings straight when doing script role binding!) Look also at
> Nenov and Dyer in IEEE ICNN 1987 San Diego conference.

It's no coincidence that symbol manipulation uses ASCII, because in a formal symbol system the shapes of the symbol tokens must be arbitrary: they cannot resemble or be in any other causal way related to the objects or states of affairs they stand for. If the symbol tokens are constrained in any way, e.g., by either resembling or being causally connected to the objects they stand for, it's another ball game. (I would add only that I still believe a significant amount of feature-learning is done during an organism's lifetime, rather than its evolutionary history -- especially in the case of human beings.)

> 4. Fodor's innateness
>
> I doubt if Fodor remembers, but he visited Yale University years ago
> (when I was a grad student there) and he gave a talk on innateness. to
> Schank's group. Well, everyone started asking about what concepts were
> innate. As I recall, we got him to admit that "telephone" was innate,
> but "Chicago" didn't seem to be innate. It was really quite amazing. It
> seemed to be the case that he postulated innateness in any case in
> which he couldn't figure out an alternative.

Since Fodor believes that the intersection of the detectable features of the members of (many? most? all?) categories is empty, he probably believes in something worse then mere innateness (for that allows for feature-detectors shaped by evolutionary "learning"). Such Cartesian or Platonic "Spandrels" would have to be the residue of what I've dubbed the "Big Bang Theory of the Origin of Knowledge." Too big a bullet for me to bite, if my thinking were to lead me inexorably in that direction...

Stevan Harnad

[Michael Dyer has responded to this; his response will be posted in the next mailing.]

[Here are Michael Dyer's Comments on Searle, which I promised to circulate. Since they are long (the whole text is around 500 lines) and quote extensively from me, I have restricted my own responses to some interpolated lines here and there, indented, and enclosed in square brackets, like this. SH]

From: Dr Michael G Dyer To: Stevan Harnad

Stevan, you wrote:

> You are apparently satisfied with the "Systems Reply." I definitely am
> not. Searle has repeatedly said that it's not that he doesn't believe
> intelligence arises from "component interactions," because he does
> believe that, for example, it arises from the component interactions in
> the brain. Nor is Searle committed to any particular kind of component
> type. He just holds (correctly, in my view) that his argument has
> shown that certain types of system (pure symbol systems, as it turns out,
> in my analysis) are the wrong types of system to exhibit intrinsic
> intentionality; their intentionality is all parasitic on our
> interpretations.

The problem with Searle's "intentionality" is that it is always fundamentally "parasitic on our interpretations" -- i.e. YOUR having intentionality (to ME) is parasitic on MY interpretation that bestows intentionality to YOU. History is full of one (often distinct racial) group denying (or reducing the amount of) intentionality "bestowed" on another group.

[No, Searle's point has has nothing to do with vague social judgments; his point is much simpler; The sentences in a book don't have intrinsic meanings. They're just scratches on paper. Their meaning depends on our interpretations. The meanings in our head do not. They are intrinsic. SH]

Maybe it depends on what you mean by "a pure symbol system". A computer is a general purpose device. It can push both numbers and symbols around since they are BOTH "symbols" (in the sense of patterns moved about and transformed in systematic ways, dependent on other patterns). There is a basic ambiguity concerning the use of the term "symbol". I myself write papers about "connectionism versus symbolism", but that discussion is at the level of what is the most natural scientific language for describing cognition, and is NOT intended to imply that somehow "connectionism" gives us some magic that cannot be done on a von Neumann machine (or Turing machine).

> [You have to distinguish between a symbolic simulation and a causal
> implementation. Some physical processes, such as transduction, can be
> simulated by symbol manipulation but they cannot be implemented as
> symbol manipulation. Whether or not connectionism is vulnerable to
> Searle's Argument depends on whether or not there is any essential
> functional or causal difference between a symbolic simulation of a net

> and a nonsymbolic implementation. SH]

I can simulate any kind of neural system (to any level of detail) you want with the 'symbols' of a von Neumann Machine (and Turing machine ultimately, since a Turing machine can simulate a Von Neumann machine, albeit incredibly more slowly). There's greater and greater cost as you try to make the simulation more detailed (e.g. if you want quarks, then I might need 1,000 supercomputers to simulate that level of detail for a microsecond), but as long as Searle accepts that it is PATTERNS of interaction, then the computer can do it, once we set up a correspondence between the patterns that the computer can realize and those patterns we are hypothesizing for whatever reality we are trying to model.

[Except for symbol manipulation itself, a symbolic simulation of a physical process is not the same as a causal implementation of that process. A simulated transducer cannot do real transduction. A simulated airplane cannot fly. Searle holds no brief against interaction patterns in general, just purely symbolic ones. SH]

> {Dyer misses] the essence of Searle's Argument: The person is not just
> being a piece of a system, HE'S DOING EVERYTHING IN THE SYSTEM.
> Whatever functions you want to point to that a symbol-manipulator is
> performing in order to pass the Turing Test, Searle can simply do them
> all himself. Introspection is the right test for whether the mental state
> that is being attributed to the system he is simulating is really there
> under conditions where he is performing every last function the system
> is performing. That's why Searle says he doesn't see much left to hang
> a mind on in the walls and ceiling of the Chinese Room and the chalk on
> the blackboard. You're tilting at windmills if you want to insist that
> THAT system has a mind. It's not that Searle is claiming NO system can
> have a mind, even one of which he is himself a component. He's just
> claiming that THAT system -- the pure symbol-cruncher under discussion
> in the Chinese Room Argument -- has no mind.

Now we really have a CLEAR DISAGREEMENT! Imagine that we encounter a strange form of intelligence from the planet XIMULA. Each XIMULAN is highly intelligent (composes operas, etc.). However, one day humans discover that the brains of the XIMULANs consist of societies of miniature creatures, called the MINIXIANS. The MINIXIANS within the brain of a given XIMULAN all work together, communicating and performing many operations, to 'control' the behavior of that XIMULAN. It turns out that you can talk to a single MINIXIAN (they're that smart). It turns out that no MINIXIAN knows really what the XIMULAN is doing (or thinking). Now we have two levels of "intentionality". You can talk to the XIMULAN and he/she will tell you about his/her loves, desires, thoughts, feelings, opera ideas, etc. But you can also talk to a given MINIXIAN about what he does and why, etc.

[This example is irrelevant because Searle would not deny it. Reread the last two sentences of the quoted passage before it. SH]

Put more simply: (1) I can use lots of smart components to make a dumb device; e.g. I can make a simple hand-calculator out of a football field of humans, all who are taught how and when to perform simple bionary operations. No human in the football field need even know what he/she is doing. (2) I can use lots of dumb components to make a smart device -- that's the goal of both AI and connectionism.

If Searle can move around fast enough (to keep the time frames in synch -- so he'd have to be VERY fast!) to perform every operation that's involved in a given program execution, then no one will be able to tell the difference. Let's assume this program's execution runs a connectionist simulation which is looking at a manufactured visual scene and answering questions about it. Now it's really Searle who is implementing the underlying machine. Now if we ask the system what it sees in the scene, it will answer. But if we ask Searle himself what he experiences he will say "Whew! I'm running around here, flipping this bit, flopping that bit, moving this electron down this data bus, ... This is really hectic!" etc. The point is, Searle's introspection on his own experience will give us NO INSIGHT into the intentionality of the connectionist system (that is talking to us about what it is seeing in the visual scene).

[The example has gotten too baroque. But if the bottom line is that this is again a case of pure symbol-manipulation and question-answering, Searle's Argument is again valid. If nonsymbolic functions such as transduction, analog transformations or essential parallelism are involved, Searle's Argument fails, as I've shown. SH]

Ever since I got into AI (and connectionism, since 1986) I have always been amazed that my students and I can design a system that, at the I/O level, will tell you, e.g. why Milton Friedman is against protectionism (e.g. see Alvarado's 1989 PhD. or Alvarado, Dyer, Flowers in AAAI-86 proceedings or in Knowledge-Based Systems (in press)), but when you open the program up, it's just the interaction of a bunch of dumb parts. If you want to make Searle do lots of LISP CARs/CDRs/CONSs (or lots of sums & sigmoidals), why should I judge the intentionality of the resulting system just because Searle's personal experience is that he's just doing lots of CDRs or sigmoidals? That would be like judging YOUR intentionality by polling one of your neurons. If I did that, I would conclude you have the same intentionality as that neuron.

[Searle's Argument, valid in my view, is an attempt to show why those pure symbol manipulation systems, ALL of whose functions he can himself perform, don't have intrinsic intentionality no matter what you can read off the top. The Argument is not valid for every possible system and every possible function. He CAN do all CARS/ CDRS, symbolic sums and symbolic sigmoid approximations. He could not do "real" multiplication, if that meant a nonlinear physical process. But then neither could a symbol-manipulator. It could just do a symbolic approximation. For implementing a mind, this kind of functional difference could be crucial. SH]

We really have to separate the business of (a) symbol grounding from that of (b) at what LEVEL of system behavior we are going to look for intentionality. It is the cheapest kind of trick for Searle to ask us to imagine doing CDRs (or neuron spiking) and then ask us to judge the intentionality of the system based on our intentionality as we do any on (or all!) of these low-level tasks.

I will tell you what the simlutaneous "doing" of all of my neurons is like -- it's ME. But just as a neuron can't tell you about ME, I cannot tell you about what it's like to be a neuron. I suspect Searle realizes that he's pulling "massive levels confusion" over on people, and if he can bolster his argument by using symbol groundings, he will, but it's a red herring.

[I must repeat, Searle has nothing against systems with dumb components, e.g., the brain; his argument only works against pure symbol systems. SH]

> And he’s simply right about that, according to me. But, as I show in my
> paper, it turns out that a very simple variant is immune to Searle’s
> objection: One in which the candidate system’s functions -- the ones
> all of which Searle is committed to performing himself, if his argument
> is to go through -- include certain nonsymbolic functions, in
> particular, sensory transduction. If, for example, the physical system
> that implements the mental state of seeing something -- that state we
> all know whether or not we’re in just as directly as we all know
> whether or not we’re understanding Chinese -- includes transducer
> function, then Searle cannot simulate it Chinese-Room style. For either
> he will have to get only the transducer’s output (in which case it’s no
> wonder he’s not seeing, since he is not performing all the functions
> the system he is simulating is supposed to perform) or he will have to
> BE the transducer, in which case he will be seeing after all.

We’re building a simulation in the UCLA AI lab that has a retina of artificial neurons. This simulation is done on a connection machine, but could be done on a von neumann machine. Given enough time, Searle himself could simulate the entire system. The system learns to describe the motion of a simulated visual object (i.e. neuron firing in a region of a 1K x 1K array of neurons making up the retina and projecting to other maps, which extract motion, direction, etc.).

[If your system uses a symbolically simulated retina, and symbolically simulated visual objects, then Searle will indeed be able to simulate it, and his Argument will apply, and he’ll be right, because the system won’t really be seeing. But if the system uses for its retina real transducers, that transduce real energy from real objects, then Searle will not be able to perform this nonsymbolic function himself, and hence whether the system could really see remains open. -- I would personally still doubt it till it could pass the Total Turing Test for some organism at least... SH]

I believe that the invariant representation (as a pattern) of, say, a cat, should be formed from visual experience in seeing the shape of cat, but that issue is not related to whether or not the whole process can be run on a computer (it can). I will send you a review I wrote of a talk Searle gave here at UCLA. His talk so infuriated me that I wrote a critique of it, but have not sent it anywhere.

[Simulated transduction is not transduction. The whole process of visual experience hence cannot be implemented on a computer, only symbolically simulated. And, as Searle showed, pure symbol systems don’t have minds. SH]

> The reason Searle’s argument works with understanding and fails with
> seeing is that the candidate system for understanding was purely
> symbolic, whereas the candidate system for sensing was not. It was this
> immunity of a seemingly trivial function (transduction) to Searle’s
> Argument that confirmed my hunch that sensory grounding may be
> critical for modeling the mind. (And of course I believe that the place
> to ground understanding is in sensory function.)

Sensory grounding is critical for modeling certain ASPECTS of mind. Grounding gives a greater richness to the symbols and allows the system to reason about 3-D objects and motions, etc. in a way that AI systems currently cannot. But we can smoothly weaken the amount of grounding. E.g.

imagine a robot/person with a lower-resolution retina, with lower resolution of sensory membranes on the hands and in the joints. Or imagine more and more "noise" in the distributed patterns that reconstruct sensory experiences. But the fact that we can build systems with symbolic systematicity properties that read fragments of editorial text (although highly primitive) indicates that those kinds of organizations capture ANOTHER ASPECT of mind, one that is only loosely coupled to grounding symbols.

[In my view, sensory grounding is not just a matter of convenience or greater "richness": it's essential for having a mind -- not for having "aspects" of a mind, but for having a mind at all. Nor do I think it's a matter of a sensory module you can simply add on to a symbol cruncher module that captures another "aspect" of mind; a grounded system is hybrid through and through. -- And I don't think having a mind is captured "aspect by aspect." I think having a mind is all or none (and that toy systems don't have any; only TTT-scale systems do). SH]

> And I think Searle would be agnostic about whether or not something had
> to be "alive" (by which I assume you mean a natural biological creature
> that is not dead) to have a mind. For example, I doubt Searle would
> have any problem with artificial neurons, as long as they had the same
> causal powers as natural neurons. But Searle's extracurricular beliefs
> are irrelevant. Insofar as they are based on the Chinese Room Argument,
> he has given strong reasons for doubting that mere symbol-crunchers
> ("Strong AI") can have minds. The rest is moot.

Searle must know that we can simulate artificial neurons on standard computers (using the same LOAD, STORE and other register operations in the machine that, by the way, are also used to simulate CDRs), so I can't believe that he would accept artificial neurons, because he can ask you to imagine being a neuron and then say "where's the intentionality?" -- THAT's why I think it boils down to Searle needing a real neuron (i.e. one that's alive). Then the whole Searlean argument will repeat itself for Artificial Life (just as it has for Searle's attack on Artificial Intelligence).

[Please reread the discussion of symbolic simulation versus causal implementation of nonsymbolic functions, e.g., transduction.]

> I repeat: although Searle hasn't said so explicitly, I will: The only
> kind of artificial system that is vulnerable to his Chinese Room
> Argument is a pure symbol manipulating system. When it comes to
> nonsymbolic functions, all bets are off. Searle mentioned in passing
> that a system with a mind must have all the causal powers of a real
> brain, but I'm sure he would agree that many of the causal powers of
> the brain may be irrelevant to its having a mind (as Stephen Hawking's
> tragic case shows so dramatically), and that all he can really insist
> on are those causal powers of the brain that are necessary and
> sufficient for having a mind (which begins to sound a little
> tautological), and that any system, natural or artificial, that had
> those, would have a mind. Apart from that, Searle simply points out,
> correctly, that the kind of system he can simulate in the Chinese room
> (a pure symbol system) clearly does not have a mind, whereas the brain
> clearly does.

Searle never said very clearly just WHAT is in those books that the person inside the chinese box (CBX) is using (in order for the CBX to speak Chinese). Here’s some content for those books: The books consist of the circuitry (as drawings, or just as symbols and pointers, whichever you prefer) of a connectionist system (with neurons as detailed as you choose). Searle grabs the chinese characters and lays them on a page that corresonds to the CBX’s "retina". Searle then traces (with many fingers, or over much time!) the values of "retinal cells" where the characters overlapped. Searle needs lots of blank pages for storing his intermediate calculations, since he has to propagate the activation values (or firing pulse patterns, if you want a more detailed level) to the next layer. Each page of the book has different layers, with directions as to which pages to flip to and what calculations to make. Some of the pages have millions/billions of patterns and these represent memories. To simulate memory recall, perhaps the CBX books direct him to do some tensor operations. In any case, after many page flippings and MANY calculations, he arrives at a pattern on a "motor page". This page explains how he should move his hand to produce chinese characters on output. Searle won’t understand what he is doing, but the CBX will perhaps be responding (in chinese) with a witty retort.

[All this symbol manipulation and symbol matching is exactly what Searle can do, and his argument correctly shows that if that’s all the system does then it does not have a mind, no matter what comes out on top. Note that connectionism is just as open to his critique as long as there’s no essential difference in causal power between a simulated and an implemented net. Note, though, that in "Minds, Machines and Searle" I give reasons why an ungrounded system would not be able to make all the right words come out on top. That success is only conjectured, after all. You may have to be able pass the TTT to pass the TT. SH]

> So the moral is simply that in the attempt isolate that subset of the
> brain’s function (or any other possible mind-supporting system’s
> function) that is necessary and sufficient for having a mind, a pure
> symbol crunching module turns out not to the right subset!

It turns out that scientists now carefully examine the CBX books and arrive at the conclusion that a subset of the patterns on the CBX pages are acting as patterns that are invariant with respect to the earlier (sensory) pages. These invariant patterns have systematic relations to one another. Some scientists decide to call these "symbols". Other scientists figure out how to simulate the entire CBX book on a von neumann machine. One crazy scientist builds a Turing machine to actually simulate the von Neumann machine, and runs the CBX program on it, and so claims that it is all "symbolic". Clearly, we have terminological problems on our hands.

[See discussion of books, simulation and implementation, passim. SH]

I said: "I’ve never heard Searle talk about the importance of grounding symbols in perception. If I’m wrong, please point me to his relevant paper(s)."

> Neither have I. But *I’ve* talked (and written) about it. And whereas I
> first thought I was disagreeing with Searle, he assured me that we were
> largely in agreement. So there you have it.

I would be worried if I found Searle agreeing with ME! :-) Seriously, it’s nice that Searle recognizes the importance of grounding. But don’t confuse that with his (false) arguments on intentionality.

> The problem is that "Gothic Style," if it’s anything at all, is an
> objective, observable property. (So is behavior under stress.) In
> principle we can all agree that any structure that has certain
> geometric properties, irrespective of any other properties, is Gothic.
> Unfortunately, this is not true of understanding (or seeing, or any of
> the other subjective properties of having a mind). For even if a system
> has all the objective, observable properties of a system that has a
> mind, it may not have a mind. And that can be true as a simple matter
> of fact, not just a metaphysical possibility, as Searle’s Chinese Room
> Argument shows.

As you can tell by now, I am totally unconvinced by Searle, so referring to his Chinese Box doesn’t help me. Is mental function the "style" of patterns, or the actual "substance" the patterns are made of? Computer scientists are now building optical computers. It will be the "style" of the photons that count for the information processing level, even if it is the "substance" of the photons that allows one to do things with greater parallelism and more rapidly.

[It’s not a matter of style vs. substance but of pure symbol manipulation versus nonsymbolic functions. SH]

> But here I actually part ways with Searle. Searle rejects all versions
> of the Turing Test, very much for the above reasons, whereas I reject
> only the traditional symbolic version of the Turing Test (symbols in,
> symbols out, and presumably just symbol-crunching in between),
> because of the Symbol Grounding Problem. The robotic version -- what
> I’ve called the Total Turing Test (TTT) -- is good enough for me,
> especially since that’s all I can ever have to go on even with my
> fellow human beings. Appearances could be deceiving there too (as the
> "Other Minds Problem" reminds us), but as far as I’m concerned, Total
> Turing power is power enough to convince me.

Turing didn’t consider direct sensory recognition and manipulation tasks in his test (TT) because he wanted to focus on language-related aspects of mind (and he didn’t want people to see the computer and make judgements based on prejudice). IF the computer in the TT does not have something like images and perceptual reasoning, then it won’t be able to hold up it’s side of a conversation about how things look and move, or about imaginary visual objects. Turing did NOT disallow the use of such conversations in his TT.

[I agree with this, in fact said it myself in "Minds, Machines and Searle," but proponents of "Strong AI" don’t. They think a repertoire of symbol strings describing visual objects would be enough.]

It is possibly the case that one could have a computer that can discuss and understand something about visual and other sensory experiences, but cannot actually move and recognize objects (in the same way that people who are learning, say, Spanish, can understand sentences they cannot generate on their own). But the extent to which you can divorce comprehension of words ABOUT the sensory world from actually patterns stored AS A RESULT OF sensory experiences is an open empirical question. I tend to agree with you that to give a system the ability to pass the TT of talking about the world will require a system that has had experience with the world (i.e. your TTT) (but mainly because forming such representations by hand will be just too hard an engineering feat --

we will want them formed automatically, from having the robot interacting with the world directly).

[I agree that there's nothing sacred about real-time history and learning; just the endstate, but I think that must must include nonsymbolic representations and capacities in order to ground the system. Real-time learning just seems the easiest and most natural way to get there from here in most cases. SH]

> Besides, no one knows what the relevant causal powers of the brain are
> anyway, so that certainly can't be the way to settle the matter. The
> brain's equivalent of your clay's "behavior under stress" may or may
> not be one of the properties any system must have in order to have a
> mind. To put it another way, it's not clear how much of the TTT will
> have to include not only the body's behavior, but the brain's. Both,
> after all, are objective, observable properties. The more you capture,
> the better your grounds for confidence. I just think (total) bodily
> performance capacity is probably a sufficient constraint.

I agree in general, but I would grant civil rights to a robot that HOBBLES and is BLIND, but can discuss sex, religion and politics, before I give civil rights to an athletic, visually acute robot with the mind of a dog.

[I'd want to protect anything that had a mind, but I happen to be a vegetarian... SH]

> Searle's "elusive" intentionality is the only kind there is. It's the
> stuff the systems that really see or feel or understand [English or
> Chinese] have (and know at first hand that they have, because they know
> exactly what it feels like to see or feel or understand [English or
> Chinese]) and that systems that only act exactly as if they see or feel
> or understand [English or Chinese] don't have (but don't know they
> don't have, because there's nobody home). It's just that I don't happen
> to believe you can get a system to ACT (TTT) exactly as if it sees...
> etc., without its really having a mind.

Searle's "intentionality" is of the worst kind! It represents a sophisticated form of prejudice. This prejudice is not based on gender, culture, or race, but on a "neural machismo-ism". I would hate to be relegated as a "non-intentional entity" by Searle just because, say, N of my neurons are damaged, so someone replaces them with a microchip that simulates their connectivity and firing patterns and then links the chip up to neighboring, "real" neurons.

[This persistent misunderstanding of Searle does not advance matters: I can only repeat: Searle would not deny the above as long as the artificial parts had the right causal powers to keep sustaining your mind. All he's claimed is that a system whose ONLY function is symbol manipulation will not have a mind. SH]

> Pure symbol-crunchers, on the other hand, cannot by their very nature
> have ALL the requisite powers, because some of the powers are
> nonsymbolic -- and, according to my theory, the symbolic ones must be
> (nonmodularly) grounded in the nonsymbolic ones; so an autonomous
> symbol crunching mudule will have no mental powers at all.

We have two extremes of a spectrum: at one end we have lots of sensory patterns and architecture for coordination and recognition tasks. At the other end we have lots of sensory-invariant patterns, related to one another in a logico-rational-linguistic manner. You're arguing, as I see it, for moving the point on this scale from the logico-linguistic extreme toward the middle and give richness to symbols. Others have argued for moving from the sensory extreme to the middle and give systematicity of abstract thought to distributed representations. They're both going to meet in the middle somewhere.

[It's not a continuum. And it's not just a matter of the old "bottom-up" approach meeting the "top-down" one in the middle: It's bottom-up all the way. The shapes of symbol tokens in formal symbol system are arbitrary. The shapes of objects in the world, and of the features in their sensory projections, are nonarbitrary. Hence a dedicated symbol system, whose functioning is constrained in this second way, is no longer a formal symbol system at all: Its primitive symbols are grounded by their resemblance and causal connection to the real physical objects they pick out and stand for. SH]

Searle, however, is on another axis: one extreme is that it's "all style", the other extreme is that it's "all substance". All AI and connectionists that I know believe that it's "all style". Every time I hear Searle, he seems to be saying it's "all substance".

[I don't think the style/substance distinction has any substance. One can be a functionalist -- distinguishing structure from multiply realizable function -- without being a symblic functionalist -- for whom all function is symbolic function, forgetting that some functions can only be simulated, but not implemented as symbol manipulation. SH]

> From a TTT standpoint, AI's feats are all mere toys.

True, but AI has excelled at combinatorial, symbolic and complex control architectures, while connectionist models have excelled at sensory processing. The grounding problem addresses the issue of just how the two approaches should be merged. Rather than a simple hybrid, I think we both feel that symbols should be embedded in, and formed out of, "connectoplasm" rather than symbols staying like ASCII and having some kind of pointer to a sensory map.

[Dunno about "connectoplasm." For me sensory transduction, sensory analogs and analog transformations are just as important; I would use the nets to learn the features of sensory categories, thereby connecting the category names -- the primitive symbols -- to the objects they refer to. SH]

I'll send you my critique of Searle's UCLA talk.

-- Michael Dyer

Free-Floating Intentionality, or, "I've Got You On My Brain" Dr Michael G Dyer wrote:

> WHY is it that you are convinced by [Searle's] argument from
> introspection (i.e. that if one imagines doing ALL of the primitive
> functions of a system, that is capable of exhibiting overall
> intelligent/intentional behavior, and one does not feel like one is

> intentional, then [it follow that] there is no intentionality)?

Because I think that the only real difference between whether a system really understands or only acts as if it understands but doesn't understand is that there is something it's like to understand, and the system that really understands has THAT, and the one that doesn't, doesn't. For example, I think that there's something it's like to understand English and something it's like to understand Chinese. I have the former and not the latter, because I understand English and not Chinese.

But before you hasten to declare that then you don't care about the difference between (1) a system that "really" understands in this introspective sense and (2) a system that only acts as if it understands, notice that you're headed for trouble there, because even YOU wouldn't be happy with (3) a system that just said "Yes, I understand" in response to everything. Why? Because you don't believe it really understands, even though it acts as if it does.

Now this was the motivation behind the original (talk-only) version of the Turing Test: How to rule out (3)? Well, require that it not be so easy to "see through" the system (as you would quickly see through the repetitious routine of (3)): Require (2) to be so hard to see through that you can't tell it apart from a real, understanding person from the way it acts (verbally); then admit you have no rational grounds for denying that it really understands. Well Searle has given you some rational grounds: Even if you can't see through (2)'s performance from the way it acts, Searle shows that (if he can himself execute ALL of its functions without himself understanding) (2) doesn't really understand because there's NOBODY HOME TO BE DOING THE UNDERSTANDING (and the only one home, Searle, doesn't understand -- if he's to be taken at his word, as I'm inclined to do; others have insisted on seeing evidence of dual personality or "speaking in tongues" in all this...).

The case to bear in mind is the hardest one: Searle has memorized all the machine tables [n.b., not IMAGINED memorizing them, but ACTUALLY memorized them; the thought experiment is imaginary, but the simulation being imagined is not imaginary]; so everything's internalized; there are no other parts, and Searle himself is all there is to the system.

Now you want to say that, despite Searle's denial, and despite the fact that there's no one else in sight, there's some understanding going on there. Well, let me ask you this then: WHO's doing the understanding? And while you're at it, where is he (she?) after the lights go out, when Searle's put away the Chinese toys and gone beddy-bye. I know that one understanding system is there, sleeping away. Do you think there are two? Since when? Since Searle memorized the tables? Talk about special creation: An understanding system comes into being because a person memorizes a bunch of meaningless symbol-manipulation rules...

> A system that has intelligence (intentionality or any other emergent
> phenomena) usually does NOT have any insight into the operations being
> formed that give rise to its emergent properties. If I had to really be
> aware of every sigmoidal function my neurons are executing I would have
> no time to be intelligent and my normal state of awareness would be
> totally bizarre. The act of doing is different from introspecting on
> the act of doing (since it is a different act of doing).

This is changing the subject. The question was only whether a system that understands Chinese has enough "insight into its operations" to know THAT it understands Chinese -- not HOW it understands Chinese, or how its neurons understand Chinese, or anything else.

> In fact, the human experience of being intentional or aware (even self-
> aware) probably RELIES on humans NOT having access to the experience of
> what it's like to do all of the underlying cognitive and/or neural
> operations that give rise to the intentionality itself (as an emergent
> phenomenon). I've tried imagining myself doing all of the LISP
> functions involved in the BORIS NLP system and I can't imagine what
> it's like to be the BORIS system. I've tried imagining doing all the
> backprop operations of a language learning system and I can't imagine
> what it's like to have it's state of consciousness (if it has one).

Unless you're a panpsychist, you don't really believe EVERYTHING has a mind. (And if you are, this discussion is pointless, because the matter's settled, there's no need to pass the TT, and even (3) understands.)

But if you're not a panpsychist then you probably better stop trying to imagine what it's like to be BORIS or backprop because, until further notice, it's probably safe to assume there's NOTHING it's like to be BORIS or backprop or most other nonliving things: Nobody home. As for your brain: You already know what it's like to be your whole brain, and perhaps even your left hemisphere plus brainstem; but hold off on what it's like to be lesser subsystems, because again, there's probably nobody home. (And as I said, lack of introspective access to the nature of the brain function underlying understanding is beside the point.)

> Let's imagine every quantum/biochemical/bioelectrical/neural operation
> done by your brain and imagine Searle imagining himself doing every
> operation (e.g. he moves ions into place for neural spiking etc.). I'm
> sure he's not going to understand what he's doing or have any insight
> into the intentionality of the brain/mind he is creating (he's just
> following a giant set of simulation instructions). So what? Who EXPECTS
> him to understand the resulting state of mind? If I talk to Searle I
> get information about what it's like to simulate a brain with some
> (unknown) intentionality. If I talk directly to the mind residing on
> that Searle-created-brain, then I can have a conversation with that
> mind and it will describe it's state of mind. If the simulation is of
> YOUR brain, then it will give me one of your favorite arguments about
> why symbol grounding is important and why Searle's chinese box is
> right. (and perhaps even ultimately be persuaded by this counter
> argument :-)

[Repeat: We're not to imagine Searle imagining, we're to imagine him DOING.]

Until further notice (i.e., until someone successfully refutes Searle's Chinese Room Argument), there's strong reason to doubt that pure symbolic simulations of either the performance of a person with a mind OR the functions of the brain have a mind. This should not be particularly surprising because, as I pointed out in "Minds, Machines and Searle," lots of symbolic simulations fail to have the critical properties of the thing they're simulating: Simulated furnaces don't heat. Simulated airplanes don't fly. Why should simulated "minds" understand? A simulated brain, after all, couldn't see, because simulated transducers don't really transduce.

I like the idea of talking to the mind "on" the "Searle-created-brain," though, especially if it's my twin. Remind me again (because I'm beginning to feel a certain familial interest), what becomes of him when Searle stops thinking about simulating my brain and goes to sleep? What's on my doppelganger's mind then, and where's his "Searle-created-brain"...?

> You seem to be thinking that, just because Searle is the only
> functioning part, there is no one else to talk to BUT Searle, but
> that's clearly NOT true! For example, when we build a natural language
> understanding system, the only functioning parts in the computer are
> the program instruction interpreter hardware circuits, bus circuits,
> memory registers, etc. But I don't even consider trying to talk to any
> of THOSE. I type input TO THE SYSTEM I have built and talk TO THAT
> SYSTEM. Even if Searle is behind every bus and register operation, I
> can still sit down and have a conversation WITH THE SYSTEM, by feeding
> input TO THE SYSTEM. It's clearly NOT Searle who replies (any more than
> it is a single register replying, who may be busy flipping its
> flip-flops, etc). -- It so happens that I can stop and ALSO talk to
> Searle, but SO WHAT?

I seem powerless to communicate to you that it's not that anyone's doubting the distinction between a system and a part of the system, or that a system can have properties its parts don't have, or that understanding systems (including the brain) can have non-understanding parts, of which one non-understanding part might even be Searle. It's still true that in the Chinese room Searle is both the part and the whole system (if not, who/what is?), so HE's the only one whose word I'm inclined to take as to whether or not he understands Chinese.

Now I realize you're going to rush to tell me that at the same time he's denying it in English, in Chinese he's telling me he CAN understand Chinese, to which I reply: Who's "he"? And where was he last Tuesday, before Searle memorized the machine table? And how come Searle doesn't know about him (he's both systems, isn't he?)? Or perhaps Searle is speaking in tongues? Or has another personality. What's his name? Shur-Lee? Ok, where were you Monday, Shur-Lee? Can anybody corroborate that, Shur-Lee?

But this is all just silly sci-fi, as far as I'm concerned, designed to milk intuitions about a counterfactual situation that's about as likely to be possible as time-travel or telepathy. I give reasons in "Minds, Machines and Searle" why no symbols-only program could ever pass the TT in the first place, so Searle would never need go to such lengths to show it had no mind -- and we need never confront paradoxes like the above. The reason is the Symbol Grounding Problem, according to which the system that can pass the (symbols-only) TT will first have to be capable of passing the TTT (Total Turing Test), which will require drawing on NONSYMBOLIC functions to implement robotic functions (like transduction) that can only be simulated, but not implemented symbolically. So Searle can't perform those functions either (and my doppelganger can forever rest in peace).

> I don't know a single computer scientist or AI researcher who does
> anything more than laugh with incredulity at Searle's argument. The
> argument is not successful in CS/AI circles and every colleague, after
> hearing Searle give a talk, is always amazed that Searle seems to
> convince non-computer scientists (who have perhaps not had the

> experience of building layers and layers of virtual machines).

And I think Searle’s Argument will go down in intellectual history as having summarily lampooned the risible credulousness with which CS/AI circles in the ’70s and ’80’s were ready to gussy up extrapolations of the performance feats of their toy symbol-crunchers with mentalistic interpretations. (It’s true that laymen have a naive blanket incredulity about artificial minds; they will stand refuted when we actually come up with one. But for now they’re on the money despite themselves. And as far as I’m concerned, pure symbol crunchers are out of the running for good.)

Stevan Harnad

Simulated Vs. Real Parallelism

To: palmer@cogsci.berkeley.edu (Stephen E. Palmer)

Steve, you wrote:

> It seems to me that one can claim that there is an important and
> principled difference between the real connectionist net (in parallel
> hardware) and the simulated one (in a standard digital computer).
> It lies not at the "computational" level of the function computed,
> but at the algorithmic level of how the computation gets done. A
> parallel algorithm is not actually the same as a simulated parallel
> algorithm at the algorithmic level because the simulated one doesn’t
> actually happen in parallel. Thus, an information processing type
> functionalist could counter what I take to be your claim that a
> functionalist would have to accept the same fate for the real net
> and the simulated net on the grounds that they differ in more than
> their mere implementation: they differ in the algorithm that is
> used to compute the function.

I know what the computational (software) and the implementational (hardware) levels are, but I’m not sure what an algorithmic "level" would be. (It seems to me that for Marr it was just a useful conceptual or methodological level for the theorist in thinking about a problem in AI.) I’m also not sure what an "information processing type functionalist" is, but let me count some of the ways you can be a "functionalist":

There are I/O functionalists, for whom I/O equivalence (Turing Test Performance), irrespective of algorithm, is fine-grained enough: Whatever passes the Chinese TT understands Chinese. These functionalists should clearly be prepared to concede themselves refuted if Searle performs all the internal functions of their candidate system and yet fails to understand.

For "algorithmic" functionalists, it would presumably be important that Searle should perform the same algorithm, and if the "algorithm" is essentially nonsymbolic -- i.e., it cannot be implemented as formal symbol manipulation on a Turing Machine -- then Searle may not be able to perform it. Those I called "essentialists" about parallelism (or about processing speed or capacity, or [sub-NP] complexity or continuity) might fall in this category -- but, as I mentioned, they would still owe us a functional reason why two algorithms that gave exactly the same results should differ so radically (one giving rise to a mind, the other not).

But would most connectionists really want to argue that a serially simulated net was using a different ALGORITHM from a net implemented in parallel? After all, Fodor & Pylyshyn have proposed that the relevant functional level for the "cognitive" functionalist is that of the "virtual" machine, and that the rest is just implementational detail. You could write and execute a parallel process on a machine that gave you virtual parallelism even though it was all implemented below the virtual level as ordinary serial computation. The "algorithms" of connectionism are presumably comprised of functions like the unit interconnectivities, the generalized delta rule (back prop), etc. If "algorithm" means anything at all, it's surely the formal rule you program your computer to follow, not the specific way the execution is implemented.

I guess it boils down to the fact that inasmuch as a parallel computation is a FORMAL notion, it must be a hardware-independent one. If a process requires the result of 30 other processes in order to be computed, it can't matter formally whether the 30 others are executed in series or in parallel, any more than it matters whether they're executed quickly or slowly (within realizable limits). And yet its this FORMAL level to which the functionalists -- I call them symbolic functionalists -- are committed.

On the other hand, as I said, to the extent that connectionists are NOT formal functionalists, they are free to claim that the implementation does matter -- but then they have the burden of showing how and why parallelism should be adequate for implementing a mind, while its functionally equivalent serial implementation is not. By way of contrast, there is no problem in accounting without any hand-waving or mysterious essentialism for the radical functional difference between my recommended candidate, transduction, and its computer-simulated counterpart.

Stevan Harnad

From: Stevan Harnad To: Michael Nitabach

Is Symbol Grounding Just Verificationism?

Mike, you wrote:

> for a symbol to be grounded is to possess a procedure for verifying on
> the basis of transducer outputs that its referent is now presenting
> itself to the organism. The only way that I see this as differing from
> standard procedural semantics is that you are postulating a new
> mechanism for obtaining these procedures--namely, as the emergent
> result of the operation of a connectionist architecture. How much
> evidence is there that *unsupervised* connectionist networks can and do
> learn appropriate input categorizations?

Some features of my model may resemble verificationist or procedural semantics. I'm not sure, because I've never seen it put together quite the way I propose. Verificationism is a philosophy, whereas what I'm addressing is the question of how to generate actual performance: discrimination, categorization and description of objects from their sensory projections. I doubt that the British Associationists or the logical positivists or the present-day verificationists have ever made this their explicit modeling task, so we cannot judge empirically whether or not they had any way to accomplish it. (The ironic thing is that even my own proposal is at a level of abstraction that is quite remote from actual implementation. Nevertheless, the constraint of aiming at implementable, performance-generating mechanisms inevitably gives rise to a different flavor of theorizing.)

I also think that, apart from proposing connectionism as the feature-learner, I've in addition proposed an original complementary role for iconic (analog) projections of the sensory surface and for context-dependent category invariants in generating discrimination, similarity judgment and identification performance. Finally, I don't see why you would think I was particularly interested in UNsupervised nets. Category learning is supervised learning par excellence, constrained by feedback from the consequences of MIScategorization. So it's the capacity of supervised nets that's at issue. (It may well prove insufficient, though.)

> How likely do you think it is that... all primitive (in the sense of
> non-decomposable) symbols can be defined solely in terms of sensory
> properties? [or] that most symbols will turn out to be complex--built
> out of a small set of elementary, sensorily grounded, ones... a browse
> through the dictionary doesn't leave the impression that many words are
> defined in terms of a small subset of sensorily grounded ones...
> (example from Fodor... "The Present Status of the Innateness Controversy".)

I think it's not only likely, but necessary. But not being a philosopher or a lexicographer, it's not "definitions" I'm looking for. It's a categorization mechanism that is capable of discriminating, naming and describing objects and states of affairs as we do. Since I don't believe in magic, and I'm certainly not prepared to subscribe to what I've dubbed the "Big Bang Theory of the Origin of Knowledge" (whereas Fodor would), according to which our categories are unlearnable from sensory data and hence must be innate, I conclude that the basis for our successful sorting and labeling must be contained in the input, otherwise we simply wouldn't be able to do it.

Fodor and many others (Wittgenstein among them) are believers in "vanishing intersections." They think that the objects and states of affairs that we name and describe DON'T HAVE the sensory invariants that are sufficient to generate our successful categorization performance (nor are they recursively grounded in the names of objects and states of affairs that do). I think they do have them, for the following reasons:

I don't think it has ever really been tested whether or not category invariants exist, because no one has really tried yet. Instead, theorists have simply entered the category hierarchy at an arbitrary point, picked a category -- say, "games," or "goodness," or "symbolist poetry" -- and declared: I can't define this in terms of features that are necessary and sufficient for being a member (and certainly not sensory ones!), therefore they do not exist. (This is discussed in the last chapter of "Categorical Perception: The Groundwork of Cognition," Cambridge University Press 1987, S. Harnad, ed.)

Well, first of all, the ontological version of this question -- the problem of defining what things really exist, in terms of their essential properties -- is simply not the psychologist's or neuroscientist's problem. We only have to explain HOW organisms discriminate, identify and describe what they CAN discriminate, identify and describe (doesn't that sound like a kind of "verificational" problem?). And I think the null hypothesis here is still that they learn to do this from input, initially sensory.

There is no "poverty of the stimulus" argument for concrete and abstract categories, as there is for natural language syntax. No one has proved or given evidence that they are unlearnable.

So I think it’s more sensible to interpret the absence of invariants as evidence of the fact that (1) invariants, like so much else in cognition, are simply inaccessible to introspection, and hence will have to be found the hard way, and that (2) the hard way (which is to model the real-time, bottom-up course of category acquisition) has simply not been explored yet!

Now I’m not saying there may not be some innate sensory categories; their sensory invariants were "learned" by evolution (and some small complement of them -- but only a very small one -- may even be like "spandrels"). But whether innate or learned, the invariants must exist, else we could not categorize successfully, as we in fact do. (Our success, however, is provisional, approximate, and dependent on the context of alternatives we’ve sampled to date; being always open to revision, it is hardly a basis for a lapidary dictionary entry.)

I am not even claiming that the primitive sensory categories -- the ones that are necessary to ground the system -- are fixed and immutable. Not only are sensory categories, like all categories, open to revision if the context of alternatives grows or changes, but a grounded system is in principle "cognitively penetrable" through and through: As long as the revision is able to percolate through the system (all of it, if necessary), there’s no obstacle to a bottom-level category being revised on the basis of top-down information -- as long as the symbols at the top are grounded! A system, once grounded, can be forever revising its foundations -- pulling itself up by the bootstraps, so to speak, as long as there’s still some provisional ground to lean on. And there always is; because once words get their meanings on the basis of approximate and provisional sensory invariants -- good enough to have picked out their referents initially, at least -- then the rest of the adjustments can all be done at the symbolic level, as long as they continue to use SOME sensory grounding.

Saul Kripke (an essentialist philosopher) inadvertently gave me a good example of this. He pointed out that if we baptized the term "gold" on these rare, yellow-colored, shiny rocks, and started to use them in trade, etc., and then we discovered that some of them were fool’s gold, which wasn’t rare or precious (or we weren’t prepared to consider it such), then of course there would be no problem about revising our features for gold so as to include gold and exclude fool’s gold (based on whatever features will reliably tell them apart). But, Kripke asked, what if it turned out that we had unknowingly baptized "gold" purely on the basis of samples of fool’s gold, never yet even having encountered real gold? Would that make any difference?

Now Kripke’s concern was with essentialism, so his punchline was that, no, it was real gold that we had intended all along. I’m not interested in essentialism; but what I conclude is that we could and would certainly keep on using the name "gold" for the real stuff, even if it turned out that what we’d been using it on until that date had all been the fake stuff. The reason is that, to an approximation, the real stuff and the fake stuff share enough sensory invariants to ground the term to begin with, and from that point on, the rest can be accomplished by revision at the symbolic level (unless the sensory invariant that does distinguish fool’s gold from real gold has no name yet; no such problem here; let’s say it was the property of having atomic number 24 [?] in the periodic table -- a perfectly well-grounded measurement, in principle).

So that’s why I say that a kind of bootstrapping and top-down-revisability exists at all levels of the system. The sensory invariants, as well as the category system as a whole, are just approximations to reality -- picking things out successfully (as dictated by feedback from the consequences of MIScategorization) on the basis of the sample of confusable alternative inputs sampled to date. No eternal "definitions" are etched in stone; just the wherewithal to keep discriminating, identifying and describing things as successfully as we do, based on the input we’ve encountered so far.

> Attempts in the Philosophy of Science to ground scientific theories
> in a small vocabulary of operational, sensory based, terms have failed
> miserably (e.g. logical positivism)... an organism’s
> conceptual armament [is] a very complex "theory" of the world...

As I said, this is not philosophy of science, it’s a model for our categorization performance capacity. And "primitives" are always revisable; they’re just needed to get the symbolic level off the ground in the first place. And "theories" don’t just float freely in the head, because of the symbol grounding problem. Their elementary terms must be connected to their (provisional) referents by some (provisionally) reliable sensory invariants. Otherwise your "theory driven" organism is just in Searle’s Chinese room of meaningless symbols.

> I don’t see how it is possible for a primitive symbol to "resemble" its
> referent, in the clear sense you have developed, while still playing
> the appropriate syntactic role in mental computation. Either an
> alternate mechanism for grounding is required, or we must develop some
> theory of cognitive processing which isn’t based on the notion of
> syntactic computation over mental representations.

As I state at the end of the paper, I think the latter is true: Since the kind of (1) bottom-up, (2) hybrid (symbolic/nonsymbolic) system required by my grounding proposal is (3) a dedicated symbol system, it’s no longer clear whether it’s really a symbol system at all, for the reason you mention (which I likewise spell out at the end of the paper): A pure symbol system manipulates symbol tokens on the basis of syntactic rules operating purely on the their (arbitrary) shapes. That’s supposed to be the only constraint. In a hybrid system grounded along the lines I propose, there is a second constraint, namely, the connection between the elementary symbols and the analog sensory projections and invariance-detectors that pick out their referents on the basis of their (nonarbitrary) shapes. With this extra constraint, all bets about formal symbol systems are off until the formal properties of such dedicated devices are investigated more deeply.

Stevan Harnad

To: Stevan Harnad From: John McCarthy

[In reply to message sent Fri, 6 Oct 89 13:45:17 EDT.]

Suppose the man in the Chinese room has memorized the procedures so he doesn’t need the book. He is then interpreting a Chinese personality. This phenomenon is is common in computer use - one program interpreting another. When it occurs, it is necessary to distinguish the information used directly by the interpreter and that used by the program being interpreted. It is necessary to distinguish both from the computer hardware.

In discussing people, it is not ordinarily necessary to distinguish between a person's body (including brain) and the personality the brain implements, because there's only one. Searle has concocted a case for which it would be necessary to distinguish between the English-speaking personality and the Chinese-speaking personality being interpreted. Since he doesn't make the distinction, he gets confused. R. L. Stevenson avoided this confusion in his "Dr. Jekyll and Mr. Hyde".

From: srh@flash.bellcore.com (stevan r harnad) To: JMC@sail.stanford.edu, srh@flash.bellcore.com

Except that I'm not prepared to believe that a person can get a case of dual personality just from memorizing a set of meaningless symbols and rules for manipulating them. Do you, really? I mean the shrinks are saying that you need stronger stuff to cause such a radical change, like early child abuse or drugs or something...

Stevan

Since Richard Yee has let the cat out of the bag with his posting (I was hoping for more replies about whether the community considered there to be any essential difference between parallel implementations and serial simulations of neural nets before I revealed why I had posted my query): I've proposed a variant of Searle's Chinese Room Argument (which in its original form I take to be decisive in showing that you can't implement a mind with a just a pure symbol manipulating system) to which nets are vulnerable only if there is no essential difference between a serial simulation and a parallel implementation. That having been said, the variant is obvious, and I leave it to you as an exercise. Here's my reply to Yee, who wrote:

> The real question that should be asked is NOT whether [Searle], in
> following the rules, understands the Chinese characters, (clearly
> he does not), but whether [Searle] would understand the Chinese
> characters if HIS NEURONS were the ones implementing the rules and he
> were experiencing the results. In other words, the rules may or may not
> DESCRIBE A PROCESS sufficient for figuring out what the Chinese
> characters mean.

This may be the real question, but it's not the one Searle's answering in the negative. In the Chinese room there's only symbol manipulation going on. No person or "system" -- no "subject" -- is understanding. This means symbol manipulation alone is not sufficient to IMPLEMENT the process of understanding, any more than it can implement the process of flying. Now whether it can DESCRIBE rather than implement it is an entirely different question. I happen to see no reason why all features of a process that was sufficient to implement understanding (in neurons, say) or flying (in real airplane parts) couldn't be successfully described and pretested through symbolic simulation. But Searle has simply shown that pure symbol manipulation ITSELF cannot be the process that will successfully implement understanding (or flying). (Ditto now for PDP systems, if parallel implementations and serial simulations are equivalent or equipotent.)

> I agree that looking at the I/O behavior outside of the room is not
> sufficient...

This seems to give up the Turing Test (Searle would shout "Bravo!"). But now Yee seems to do an about-face in the direction of resurrecting the strained efforts of the AI community to show that formal symbol manipulating rule systems have not just form but content after all, and CAN understand:

> The determination of outputs is under the complete control of the
> rules, not [Searle]. [Searle] has no free will on this point (as he
> does in answering English inputs)... although it is clearly true that
> (Chinese characters) have form but absolutely no content for
> [Searle]... [w]hether or not the *content* of this symbol is recognized,
> is determined by the rules... the Chinese symbols were indeed correctly
> recognized for their CONTENT, and this happened WITHIN the room...
> the process of understanding Chinese is [indeed] occurring.

NB: No longer described or simulated, as above, but actually OCCURRING. I ask only: Where/what are these putative contents (I see only formal symbols); and who/what is the subject of this putative understanding (I see only Searle), and would he/she/it care to join in this discussion?

Now in my case this glibness is really a reflection of my belief that the Turing Test couldn't be successfully passed by a pure symbol manipulator in the first place (and hence that this whole sci-fi scenario is just a counterfactual fantasy) because of the symbol grounding problem. But Yee -- though skeptical about the Turing Test and seemingly acknowledging the simulation/implemetation distinction -- does not seem to be entirely of one mind on this matter...

> [The problem is] a failure to distinguish between a generic Turing
> Machine (TM) and one that is programmable, a Universal Turing Machine
> (UTM)... If T, as a parameter of U, is held constant, then y = T(x) =
> U(x), but this still doesn't mean that U "experiences x" the same way T
> does. The rules that the person is following are, in fact, a program
> for Chinese I/O... I take my own understanding of English (and a little
> Chinese) as an existence proof that [Understanding is Computable]

"Cogito Ergo Sum T"? -- Descartes would doubt it... I don't know what Yee means by a "T," but if it's just a pure symbol-cruncher, Searle has shown that it does not cogitate (or "experience"). If T's something more than a pure symbol-cruncher, all bets are off, and you've changed the subject.

Stevan Harnad

References:

Searle, J. (1980) Minds, Brains and Programs. Behavioral and Brain Sciences 3: 417 - 457.

Harnad, S. (1989) Minds, Machines and Searle. Journal of Experimental and Theoretical Artificial Intelligence 1: 5 - 25.

Harnad, S. (1990) The Symbol Grounding Problem. Physica D (in press)

To: delta@csl36h.ncsu.edu (Thomas Hildebrandt)

Tom, thanks for the contribution to the parallel/serial discussion; I understand your view is that there's no difference but speed.

You added:

> I notice that the definition of 'meaning' and 'understanding' are
> absent. Whatever 'understanding' is supposed to be, I say that
> technically it does not exist. Everything that the human mind does, it
> does algorithmically.

You mentioned you didn't know Searle's argument, so that might be the reason for the misunderstanding: You don't need to define "meaning" or "understanding." You just need to understand this: Searle does understand (whatever that means) English and he does not understand Chinese. He knows it, because he knows what it's like to understand English and he doesn't have that for Chinese. (If you say you don't understand what he means by that, we can't go on with this discussion. I certainly understand what he means by that distinction.)

Now Searle's argument is that if he did everything the symbol manipulator that could passed the Turing Test in Chinese did -- memorized its symbolis machine table, performed all the symbol manipulations himself -- he would STILL not understand Chinese (even though he was following the symbolic algorithm), therefore no system that did only that could understand anything either.

Your claim that to understand is just to follow an algorithm does not capture that distinction, and the distinction is quite basic and real; it's at the heart of the mind-body problem. It does not depend on commitment to any prior definition of what understanding is. On the contrary, your reply is the one that tries to settle an empirical difference by a definition (and in the process, loses the difference altogether).

Here are the self-explanatory abstracts of two papers I've written on the matter.

Stevan Harnad

-----

(1) Minds, Machines and Searle J Exper. Theor. A.I. 1(1) 1989

Stevan Harnad Department of Psychology Princeton University Princeton NJ 08544

Searle's celebrated Chinese Room Argument has shaken the foundations of Artificial Intelligence. Many refutations have been attempted, but none seem convincing. This paper is an attempt to sort out explicitly the assumptions and the logical, methodological and empirical points of disagreement. Searle is shown to have underestimated some features of computer modeling, but the heart of the issue turns out to be an empirical question about the scope and limits of the purely symbolic (computational) model of the mind. Nonsymbolic modeling turns out to be immune to the Chinese

Room Argument. The issues discussed include the Total Turing Test, modularity, neural modeling, robotics, causality and the symbol-grounding problem.

Summary and Conclusions

Searle's provocative "Chinese Room Argument" attempted to show that the goals of "Strong AI" are unrealizable. Proponents of Strong AI are supposed to believe that (i) the mind is a computer program, (ii) the brain is irrelevant, and (iii) the Turing Test is decisive. Searle's argument is that since the programmed symbol-manipulating instructions of a computer capable of passing the Turing Test for understanding Chinese could always be performed instead by a person who could not understand Chinese, the computer can hardly be said to understand Chinese. Such "simulated" understanding, Searle argues, is not the same as real understanding, which can only be accomplished by something that "duplicates" the "causal powers" of the brain. In the present paper the following points have been made:

(1) Simulation versus Implementation:

Searle fails to distinguish between the simulation of a mechanism, which is only the formal testing of a theory, and the implementation of a mechanism, which does duplicate causal powers. Searle's "simulation" only simulates simulation rather than implementation. It can no more be expected to understand than a simulated airplane can be expected to fly. Nevertheless, a successful simulation must capture formally all the relevant functional properties of a successful implementation.

(2) Theory-Testing versus Turing-Testing:

Searle's argument conflates theory-testing and Turing-Testing. Computer simulations formally encode and test models for human perceptuomotor and cognitive performance capacities; they are the medium in which the empirical and theoretical work is done. The Turing Test is an informal and open-ended test of whether or not people can discriminate the performance of the implemented simulation from that of a real human being. In a sense, we are Turing-Testing one another all the time, in our everyday solutions to the "other minds" problem.

(3) The Convergence Argument:

Searle fails to take underdetermination into account. All scientific theories are underdetermined by their data; i.e., the data are compatible with more than one theory. But as the data domain grows, the degrees of freedom for alternative (equiparametric) theories shrink. This "convergence" constraint applies to AI's "toy" linguistic and robotic models too, as they approach the capacity to pass the Total (asymptotic) Turing Test. Toy models are not modules.

(4) Brain Modeling versus Mind Modeling:

Searle also fails to appreciate that the brain itself can be understood only through theoretical modeling, and that the boundary between brain performance and body performance becomes arbitrary as one converges on an asymptotic model of total human performance capacity.

(5) The Modularity Assumption:

Searle implicitly adopts a strong, untested "modularity" assumption to the effect that certain functional parts of human cognitive performance capacity (such as language) can be be successfully modeled independently of the rest (such as perceptuomotor or "robotic" capacity). This assumption may be false for models approaching the power and generality needed to pass the Turing Test.

(6) The Teletype Turing Test versus the Robot Turing Test:

Foundational issues in cognitive science depend critically on the truth or falsity of such modularity assumptions. For example, the "teletype" (linguistic) version of the Turing Test could in principle (though not necessarily in practice) be implemented by formal symbol-manipulation alone (symbols in, symbols out), whereas the robot version necessarily calls for full causal powers of interaction with the outside world (seeing, doing AND linguistic competence).

(7) The Transducer/Effector Argument:

Prior "robot" replies to Searle have not been principled ones. They have added on robotic requirements as an arbitrary extra constraint. A principled "transducer/effector" counterargument, however, can be based on the logical fact that transduction is necessarily nonsymbolic, drawing on analog and analog-to-digital functions that can only be simulated, but not implemented, symbolically.

(8) Robotics and Causality:

Searle's argument hence fails logically for the robot version of the Turing Test, for in simulating it he would either have to USE its transducers and effectors (in which case he would not be simulating all of its functions) or he would have to BE its transducers and effectors, in which case he would indeed be duplicating their causal powers (of seeing and doing).

(9) Symbolic Functionalism versus Robotic Functionalism:

If symbol-manipulation ("symbolic functionalism") cannot in principle accomplish the functions of the transducer and effector surfaces, then there is no reason why every function in between has to be symbolic either. Nonsymbolic function may be essential to implementing minds and may be a crucial constituent of the functional substrate of mental states ("robotic functionalism"): In order to work as hypothesized (i.e., to be able to pass the Turing Test), the functionalist "brain-in-a-vat" may have to be more than just an isolated symbolic "understanding" module -- perhaps even hybrid analog/symbolic all the way through, as the real brain is, with the symbols "grounded" bottom-up in nonsymbolic representations.

(10) "Strong" versus "Weak" AI:

Finally, it is not at all clear that Searle's "Strong AI"/"Weak AI" distinction captures all the possibilities, or is even representative of the views of most cognitive scientists. Much of AI is in any case concerned with making machines do intelligent things rather than with modeling the mind.

Hence, most of Searle's argument turns out to rest on unanswered questions about the modularity of language and the scope and limits of the symbolic approach to modeling cognition. If the modularity assumption turns out to be false, then a top-down symbol-manipulative approach to explaining the mind may be completely misguided because its symbols (and their interpretations) remain ungrounded -- not for Searle's reasons (since Searle's argument shares the cognitive modularity assumption with "Strong AI"), but because of the transdsucer/effector argument (and its ramifications for the kind of hybrid, bottom-up processing that may then turn out to be optimal, or even essential, in between transducers and effectors). What is undeniable is that a successful theory of cognition will have to be computable (simulable), if not exclusively computational (symbol-manipulative). Perhaps this is what Searle means (or ought to mean) by "Weak AI."

---

(2) THE SYMBOL GROUNDING PROBLEM [To appear in Physica D, 1990]

ABSTRACT: There has been much discussion recently about the scope and limits of purely symbolic models of the mind and about the proper role of connectionism in cognitive modeling. This paper describes the "symbol grounding problem" for a semantically interpretable symbol system: How can its semantic interpretation be made intrinsic to the symbol system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols? The problem is analogous to trying to learn Chinese from a Chinese/Chinese dictionary alone.

A candidate solution is sketched: Symbolic representations must be grounded bottom-up in nonsymbolic representations of two kinds: (1) iconic representations, which are analogs of the proximal sensory projections of distal objects and events, and (2) categorical representations, which are learned and innate feature-detectors that pick out the invariant features of object and event categories from their sensory projections. Elementary symbols are the names of these object and event categories, assigned on the basis of their (nonsymbolic) categorical representations. Higher-order (3) symbolic representations, grounded in these elementary symbols, consist of symbol strings describing category membership relations ("An X is a Y that is Z").

Connectionism is one natural candidate for the mechanism that learns the invariant features underlying categorical representations, thereby connecting names to the proximal projections of the distal objects they stand for. In this way connectionism can be seen as a complementary component in a hybrid nonsymbolic/symbolic model of the mind, rather than a rival to purely symbolic modeling. Such a hybrid model would not have an autonomous symbolic "module," however; the symbolic functions would emerge as an intrinsically "dedicated" symbol system as a consequence of the bottom-up grounding of categories' names in their sensory representations. Symbol manipulation would be governed not just by the arbitrary shapes of the symbol tokens, but by the nonarbitrary shapes of the icons and category invariants in which they are grounded.

J.H. Hexter and the Road to Beijing

To: Steve Smoliar From: Stevan Harnad

Steve, you ask:

> What can a memory be OF, if not an experience?

It can be of how to do something, what something means, what something looks/sounds like (in general, not on a particular occasion), etc.

> I, too, know that I understand English and I don't understand Chinese;
> but what do I know about YOU... all I know about your understanding
> of Chinese is what you have told me, which is neither better nor worse
> than a computer which is programmed to respond "Yes" when I type in "Do
> you understand me?"

Oh, my! Do things ever manage to get garbled up! Whether what I have told you is better or worse than what a computer tells you is what is AT ISSUE here. You can't trot it out as a first principle. And you know a whale of a lot more about me than the Turing Test [TT] (especially now that we've met): You know I'm a human being, made out of the same wetware as you. So you have much better antecedent grounds for trusting me than your VAX if I say I understand English and not Chinese. Ditto for Searle. And ditto for Searle when he's doing everything the VAX is doing when you think it's understanding Chinese, and he is telling you he DOESN'T understand Chinese, and that the Chinese symbols he's manipulating for you are just gibberish to him.

> In order for me to form any knowledgeable opinion of your understanding
> of Chinese, I would probably have to observe you in a Chinese environment.

Now this is no longer just the TT (symbols in, symbols out); it's the Total Turing Test [TTT], calling for full sensorimotor interaction with the world. And for a system to pass that, it must have nonsymbolic functions, transduction at the least. Transduction can be simulated, but not implemented symbolically. Hence it is immune to Searle's Chinese Room Argument. That's what symbol grounding is about. I've repeated this many times before. You should give some sign of having understood it rather than repeating the step that preceded it every time...

> does Diana Rigg know if you understand Chinese?... she is an actress
> and therefore used to dealing with people playing roles in which they
> say things which are not true about themselves.

Indeed she is. And she presumably also knows what it's like to play a role in a language she doesn't understand...

> So, if she REALLY had to know if you understood Chinese... She
> would.... put you in an environment in which your knowledge of Chinese
> could be observed... In other words she would end up testing you
> exactly the same way we test the man in Searle's room and use those
> interactions to say whether or not you understood Chinese.

This mixing up of issues is getting to be too much for me: Searle's room is just the TT (symbols in, symbols out), NOT the TTT. That's the basic point of the symbol grounding problem. It NECESSARILY takes more to pass the TTT than the TT (in particular, it REQUIRES nonsymbolic functions). And, as I've said over and over, I think even the conjecture that the TT could be successfully passed by a pure symbol cruncher is counterfactual, again because of the symbol grounding problem. I'm sure it would require full TTT power to pass the TT, and that in turn would be immune to the Chinese Room Argument.

> I think the reason many of us in AI do not take Searle's argument very
> seriously is that... it really does not tell us anything we want to
> know. It certainly does not tell us very much about how intelligent
> agents get on in the world by interacting with other intelligent
> agents. Ultimately, it seems to be yet another attempt to demonstrate
> that there is something in being human which rises above the capability
> of any machine. Minsky, on the other hand, wishes to assure us that
> there is no stigma in being perceived as a machine, since machines have
> the power to behave with as much complexity as humans.. if not more
> so. Is this not a healthier attitude to assume if one is interested in
> the workings of the mind?

It's not a matter of "attitude" (which is ALL Minsky is giving us any more), or of "complexity." Searle doesn't show us which way is right, but he sure does show us which way is wrong: Pure symbol manipulation. Stereotyping him as a Luddite is just obfuscation. Searle repeats over and over that he knows perfectly well we're machines; it's a question of what KIND of machine we are. And one kind we're NOT is pure symbol crunchers. And that's good to know. I feel it's time for me to trot out my J.H. Hexter quote again (always at this same juncture):

in an academic generation a little overaddicted to "politesse," it may be worth saying that violent destruction is not necessarily worthless and futile. Even though it leaves doubt about the right road for London, it helps if someone rips up, however violently, a 'To London' sign on the Dover cliffs pointing south...

Steve, I'm afraid I can't discuss this further until I hear a new argument, or at least some sign of someone's having understood SOME portion of the points I've been making.

Stevan Harnad

To: P & P Churchland From: Stevan Harnad THINKING THE UNTHINKABLE, OR, RUN THAT BY ME AGAIN, FASTER

Hi Pattie and Paul:

Thanks for sending me your Scientific American draft. I've seen Searle's companion draft too. Here are a few comments:

(1) Unfortunately, in suggesting that Searle is the one who is begging the question or assuming the truth of the hypothesis that syntax alone can consititute semantics you seem to have the logic reversed: In fact, Searle's the one who's TESTING that hypothesis and answering that question; and the Chinese-Room thought-experiment shows that the hypothesis fails the test and the answer is no! It is the proponents of the "systems reply" -- which merely amounts to a reiteration of the hypothesis in the face of Searle's negative evidence -- who are begging the question.

By the way, in endorsing the systems reply in principle, as you do (apparently only because of its counterintuitiveness, and the fact that other counterintuitive things have, in the history of science, turned out to be correct after all), you leave out Searle's very apt RESPONSE to the counterintuitive idea that the "system" consisting of him plus the room and its contents might still be understanding even if he himself is not: He memorizes the rules, and henceforth he IS all there is to the system, yet still he doesn't understand Chinese. (And I hope you won't rejoin with the naive hackers' multiple-personality gambit at this point, which is CLEARLY wanting to save the original hypothesis at any counterfactual price: There is no reason whatsoever to believe that simply memorizing a bunch of symbols and symbol manipulation rules and then executing them is one of the etiologies of multiple personality disorder!)

As to the speed factor: Yes, that is one last holdout, if it is in fact true that Searle could never pass the Chinese TT in real time. But that's at the price of being prepared to believe that the difference between having and not having a mind is purely a function of speed! The phenomenon of phase transitions in physics notwithstanding, that sounds like a fast one, too fast for me to swallow, at any rate. Besides, once he's memorized the rules (PLEASE don't parallel the speed argument with a capacity argument too!), it's not clear that Searle could not manage a good bit of the symbol manipulation in real time anyway. The whole point of this exercise, after all, is to show that thinking can't be just symbol manipulation -- at ANY speed.

I don't know about you, but I've never been at all happy with naive hackers' claims that all there is to mind is the usual stuff, but (1) faster, (2) bigger, and (3) more "complex." I think the real punchline's going to turn out to be a good bit more substantive than this hand-waving about just (a lot) more of the same...

(2) All your examples about the groundlessness of prior skepticism in the face of physical theories of sound, light and life were (perhaps unknowingly) parasitic on subjectivity. Only now, in mind-modeling, is the same old problem finally being confronted on its home turf. But All prior bets are off, since those were all away-games. The buck, as Tom Nagel notes, stops with qualia. I'll show this specifically with the example below.

(3) It is ironic that your example of light = oscillating electromagnetic radiation should also hinge on speed (frequency). You say that Searle, in a parallel "simulation," would be waving the magnet much too slowly, and would then unjustly proclaim "Luminous room, my foot, Mr. Maxwell. It's pitch black in here!" But here you see how all these supposedly analogous forms of skepticism are actually parasitic on subjectivity (with shades of Locke's primary and secondary qualities): Because of course the only thing missing is the VISIBILITY of light at the slow frequency. It made perfect sense, and could have been pointed out all along, that, if fast electromagnetic oscillations really are light, then it might only be visible to the eye in some of its frequency ranges, and invisible but detectable by other instruments in other frequency ranges.

That story is perfectly tenable, and in no way analogous to Searle's Argument, because it is objective: It's not "what it's like to see light" (a subjective, "secondary" quality) that the Maxwellian equation of light with EM radiation is trying to explain, it's the objective physical property that, among other things, happens to be the normal cause of the subjective quality of seeing light. The mistake the sceptics were making is clearly NOT the same as Searle's. They were denying an objective-to-objective equation: One set of objective physical properties (among them the power to cause us to see light) was simply being shown to be the same as another set of objective physical properties. No one was trying to equate THE SUBJECTIVE QUALITY OF LIGHT ITSELF with something objective. (Not until lately, that is.)

So, whereas concerns about subjectivity might indeed have been the source of the earlier scepticism, all that scepticism was simply misplaced. It was much ado about nothing. Ditto with sound and life: Subjectivity, though lurking in each case, was really never at issue. As Nagel puts it, one set of appearances was simply being replaced by (or eliminated in favor of) another, in the new view, however surprising the new appearances might have appeared. But no one was really trying to replace APPEARANCES THEMSELVES by something else, by the stuff of which all appearances would then allegedly be made: THAT would have been a harder nut to crack.

But that's the one modern mind-modeling is up against, and Nagel is right that this is another ball game altogether (my "home-game" analogy was an understatement -- and the metaphors are as mixed as nuts by now...). So no analogies carry over. It's not that the rules have changed. It's just that nothing remotely like this has ever come up before. So, in particular, you are NOT entitled to help yourself to the speed analogy in trying to refute Searle's Chinese Room Argument. Because whereas it would have been no problem at all for Maxwell to "bite the bullet" and claim that very slow EM oscillation was still light, only it wasn't visible, one CANNOT say that very slow symbol-manipulation is still thinking only it's... what? "Unthinkable?" You took the words out of my mouth.

(4) Your point about the immunity of parallel processes to the Chinese Room Argument (unlike similar points about speed, capacity or complexity) has somewhat more prima facie force because it really is based on something Searle can't take over all by himself, the way he could with symbol manipulation. On the face of it, Searle couldn't BE the parallel system that was passing the TT in Chinese in the same way he could BE the serial symbol system, so he could not take the next step and show that he would not be understanding Chinese if he were (and hence that neither would the system he was duplicating).

This is why I suggested to Searle that his "Chinese Gym" Argument fails to have the force of his original Chinese Room Argument, and is indeed vulnerable to a "systems reply." It's also why I suggested the "three-room" argument to Searle (which he has now expanded to four rooms in his Scientific American piece, somewhat unnecessarily, in my view, to accommodate the Chinese Gym too), which is completely in the spirit of the original Chinese Room Argument and puts the burden of evidence or argument on the essential parallelist, where it belongs. Here is the critical excerpt from my comments on an earlier draft by Searle:

> So I respectfully recommend that you jettison the Chinese Gym Argument
> and instead deal with connectionism by turning the Chinese Room
> Argument on its head, as follows. Suppose there are three rooms:
>
> (1) In one there is a real Net (implemented as physical units, with
> real physical links, real excitatory/inhibitory interconnections
> real parallel distributed processing, real backpropping, etc.) that
> could pass the Turing Test in Chinese (Chinese symbols in, Chinese
> symbols out).
>
> (2) In the second there is a computer simulation of (1) that likewise
> passes the TT in Chinese.
>
> (3) In the third is Searle, performing ALL the functions of (2),
> likewise passing the Chinese TT (while still not understanding, of

> course).
>
> Now the connectionists have only two choices:
>
> Either they must claim that all three understand Chinese (in which case
> they are back up against the old Chinese Room Argument), or the
> essentialists among them will have to claim that (1) understands but (2)
> and (3) do not -- but without being able to give any functional reason
> whatsoever why.

So this is what parallelism is up against. I also went on to query the Connectionists on this, as follows (and received multiple replies, most along the lines of the 1st two, which I include below):

> From: Stevan Harnad
> To: connectionists@cs.cmu.edu
> Subject: Parallelism, Real vs. Simulated: A Query
>
> "I have a simple question: What capabilities of PDP systems do and
> do not depend on the net's actually being implemented in parallel,
> rather than just being serially simulated? Is it only speed and
> capacity parameters, or something more?"
> ---

>
> (1) From: skrzypek@CS.UCLA.EDU (Dr. Josef Skrzypek)
> Cc: connectionists@cs.cmu.edu
>
> Good (and dangerous) question. Applicable to Neural Nets in general
> and not only to PDP.
>
> It appears that you can simulate anything that you wish. In principle
> you trade computation in space for computation in time. If you can
> make your time-slices small enough and complete all of the necessary
> computation within each slice there seem to be no reason to have
> neural networks. In reality, simulation of synchronized, temporal
> events taking place in a 3D network that allows for feedback pathways
> is rather cumbersome.
>
> (2) From: Michael Witbrock
>
> I believe that none of the properties depend on parallel implementation.
>
> There is a proof of the formal equivalence of continuous and discrete
> finite state automata, which I believe could be transformed to prove the
> formal equivalence of parallel and serially simulated pdp models.

Except for some equivocal stuff on "asynchronous" vs "synchronous" processes, about which some claimed one thing and others claimed the opposite, most respondents agreed that the parallel and serial implementations were equivalent. Hence it is not true, as you write, that parallel systems are "not threatened by [Searle's] Chinese Room argument." They are, although someone may still come up with a plausible reason why, although the computational difference is nonexistent, the implementational difference is an essential one.

And that may, logically speaking, turn out to be (one of the) answer(s) to the question of which of the "causal powers" of the brain are actually relevant (and necessary/sufficient) for producing a mind. I think Searle's Argument (and my Symbol Grounding Problem) have effectively put pure symbol manipulation out of contention. I don't think "the same, only faster, bigger, or more complex" holds much hope either. And parallelism stands a chance only if someone can show what it is about its implementation in that form, rather than in fast serial symbolic form, is critical. My own favored candidate for the "relevant" property, however, namely, sensory grounding, and sensory transduction in particular, has the virtue of not only being, like parallelism, invulnerable to the Chinese Argument (as I showed in "Minds, Machines and Searle"), but also being a natural candidate for a solution to the Symbol Grounding Problem, thereby, unlike paralellism, wearing the reason WHY it's critical on its sleeve, so to speak.

(5) Finally, you write "We, and Searle, reject the Turing Test as a sufficient condition for conscious intelligence." In this I must disagree with both (or rather, all three) of you: The logic goes like this. So far, only pure symbol crunching has been disqualified as a candidate for being the sufficient condition for having a mind. But don't forget that it was only a conjecture (and in my mind always a counterfactual one) that the standard (language-only) Turing Test (only symbols in, and symbols out), the TT, could be successfully passed by a pure symbol cruncher. Searle's argument shows that IF the TT could be passed by symbol crunching alone, THEN, because of the Chinese Room Argument, it would not have a mind, and hence the TT is to be rejected.

Another possibility remains, however, which is that it is impossible to successfully pass the TT with symbol crunching alone. The truth may instead be that any candidate that could pass the TT would already have to have and draw upon the causal power to pass the TTT, the Total Turing Test, which includes all of our robotic, sensorimotor capacities in the real world of objects. Now the TTT necessarily depends on transduction, which is naturally and transparently immune to Searle's Chinese Room Argument. Hence there is no reason to reject the TTT (indeed, I would argue, there's no alternative to the TTT, which, perhaps expanded to include neural function -- the "TTTT"? -- is simply equivalent to empiricism!). And if a necessary condition for passing the TT is the causal power to pass the TTT, then there's really no reason left for rejecting the TT either.

Best wishes,

Stevan Harnad

[Recipients of the Searle/symbol-grounding discussions: If you want your name removed from this discussion list, please let me know. -- SH]

TWO WAYS TO REFUTE SEARLE

John McCarthy writes:

> To people in AI, obtuse philosophers like Searle and Harnad are too
> long winded and obscure to be worth trying to decipher. This is
> especially true, because it seems they aren't trying to raise any
> practical difficulties that would have to be overcome in order to make
> useful systems at any level of performance. They merely say that no
> level of performance would count as intelligent.

It might be a good idea to try to decipher the arguments anyway, just in case there is something practical there, such as (1) how pure symbol manipulation cannot be grounded in the objects it can be interpreted as representing, and hence cannot be the right model for mental activity, which IS grounded in the objects it represents; and (2) how successful performance on the TTT would count as intelligent, but because of (1), pure symbol manipulation will be unable to achieve it.

[Ceterum sentio: I am not a philosopher but a psychologist. I also happen to be a critic of Searle's; but because I nevertheless refuse to make common cause with fallacious counterarguments, I invariably spend more air time debunking wrong-headed criticisms of Searle than in discussing my own differences with him. See "Minds, Machines and Searle," J. Exp. Theor. A.I. 1(1) 1989.]

> If we avoid these distractions for now, there will be more hope of an
> AI program during their lifetimes that will overwhelm their arguments.
> It should be capable of following the most obscure and lengthy
> arguments and refuting them at any desired length, from a paragraph to
> a three volume treatise.

To write a symbol manipulating program that can attempt to counter my arguments, as AI researchers have attempted to counter them, i.e., unsuccessfully, is surely feasible, but would only amount to yet another arbitrary toy fragment of the TT (not to be confused with the TTT). I'll settle for either a system that can pass the TTT (which, as I showed in "Minds, Machines, and Searle," would be automatically immune to the Chinese Room Argument) -- or just a successful refutation by a human being.

> Maybe for debugging it, we will also need a program capable
> of generating the arguments.

For that, the arguments will have to be deciphered and understood first.

> However, perhaps I have them wrong. Do Searle and Harnad claim to have
> discovered any difficulty in achieving any particular behavioral
> performance? If so, what is the simplest thing they imagine we can't
> program computers to do that humans can do?

Arbitrary performance fragments won't do, because we're not just looking for more clever toys, but the real thing; nothing less than TTT performance will do the job.

Stevan Harnad

ON EQUIVALENCE: WEAK, STRONG and IMPLEMENTATIONAL

AMR@ibm.com Alexis Manaster Ramer IBM TJW Research Center wrote:

> The correct computational position is NOT the old-fashioned "strong AI"
> position, but rather it holds that just because a program or TM or
> grammar passes the TT, it does NOT mean that it possesses these
> properties, simply because these properties might attach to some
> structural properties of the biological machines that the conventional
> programs, TMs, and grammars do not happen to share.

Advocates of the "systems reply" say the "system" must understand, even though Searle doesn’t, because "it" is passing the TT. If the computational position you describe is correct, they are wrong, no one is understanding, even though the TT is being passed. Searle would be pleased. But what about the computer that Searle is imitating? Is IT understanding?

First of all, I’m afraid "strong" and "weak" equivalence is not enough for the distinction you are making here. Strong equivalence, unless I am radically mistaken, refers to ALGORITHMS, not to their implementation. So two different (strongly nonequivalent) algorithms may be I/O equivalent (weakly equivalent): They may generate exactly the same performance, but by different means. How may the means differ? Algorithms (including numerical ones) are, presumably, symbolic: They are formal syntactic rules for manipulating symbols according to their (arbitrary) shapes. (I will return to the special cases of parallel and analog computation later.) So weakly equivalent algorithms achieve the same results by using different symbol manipulation rules.

Implementation is another matter. ALL functionalists, computationalists among them, agree that the algorithmic or computational or software level of function is independent of the implementational or hardware level. In other words, two different implementations of the SAME algorithm would be strongly equivalent even though they are implementationally nonequivalent.

Now Searle’s Chinese Room Argument is not only decisive against weakly equivalent symbolic manipulation rules; it’s also decisive against strongly equivalent ones! But if someone wants to go even further and say that the same algorithm implemented one way does understand and implemented another way does not, who is to decide which implementation is OK? Maybe I like my VAX and you like your SUN, so I claim my program understands on my VAX but not on your SUN? Mere implementational differences are arbitrary unless you can pinpoint what it is about the implementational difference that makes the relevant mental difference. In any case, you’re right, if implementational differences are decisive, then the VAX could understand even though Searle doesn’t, even though they’re performing exactly the same algorithm. The question still remains, though, why should I then believe the VAX understands when it does the same thing Searle does, without understanding?

Parallel and continuous processes cannot be implemented in a serial symbol manipulator (a Turing Machine or a von-Neumann-style digital computer), so implementational considerations there are special. But since they can be approximated arbitrary closely, the burden is on the implementationalist to show what is special about the nonsymbolic implementation that makes it alone mind-supporting.

My own favored candidate for an essentially nonsymbolic function that may be crucial to implementing a mind is transduction, which can only be simulated symbolically, in the way flying can be simulated but not implemented sybolically: Simulated flying is not an APPROXIMATION to flying: It's no kind of flying at all. Hence if transduction is essential to mental function, mental function cannot be purely symbolic. Transduction is also (1) immune to Searle's Chinese Room Argument (he cannot implement transduction except by BEING a transducer) and (2) a (partial) solution to the symbol grounding problem. (All this is described in "Minds, Machines and Searle.")

> the theory of complexity focuses a lot of attention on the distinction
> (which is NOT discernable by the TT) between deterministic and
> nondeterministic machines

To the extent that statistical algorithms are at issue, we're still talking about symbol manipulation. As to the implementation of truly nondeterministic processes (quantum computers, or what have you), we will have to see what specific role is to be performed by them, and how and why, before we can decide whether they are the crucial function for implementing a mind.

> two grammars which generate the same language may differ in strong
> generative capacity... [This explains] why a grammar A is not a good
> model of the language faculty for some language and a grammar B is
> (even though both generate the language).

Irrelevant to Searle. See strong vs. weak equivalence, above.

> It remains to specify (a) what kinds of differences may count as
> differences of strong generative capacity and (b) how it is possible to
> deduce which of two weakly equivalent systems we are dealing with (and
> a fortiori to tell which is, say, conscious and which is not). It seems
> to me that the answer to (a) is that any differences at all can
> conceivably be relevant. Thus, complexity theory has focused on certain
> rather abstract ones, notably determinism and nondeterminism, but given
> two physically realized systems, e.g. a PC and an AT (even if they are
> running what is conventionally considered to be the same program), we
> may insist that for some purpose we need to pay attention to the
> distinct ions inherent in their use of different chips. By the same
> reasoning, if we need to, we can get even lower into the chemical and
> physical structure of the two systems. Thus, it would in principle be
> entirely consistent with the computational view of things if we
> discovered that some particular property of living beings depended on
> their being made up of protein etc., for example.

This is all MUCH too Searlian for me! He would agree with much more of it than I would. I think that the other-minds problem makes it impossible to "deduce" who's conscious other than yourself. I think that to pass the standard (linguistic) Turing Test (TT) (symbols in, symbols out) you need the full causal wherewithal to pass the Total (robotic) Turing Test (TTT) (our full sensorimotor interaction with the world of objects and people); perhaps the fine-tuning will come from the TTTT (our neural "performance" too), but that's as far as it goes. With the TTTT we've come to the limits of our empirical resources. I'm still the kind of functionalist who thinks the TTT is where the real action is, and the rest is just implementational details. But rather than being merely a symbolic functionalist, I'm a robotic functionalist: Nonsymbolic function is function too, and I think it's critical for grounding symbolic function.

> There exist TMs, grammars, programs which have the exact same I/O
> behavior as any given biological system [but it would be] extremely
> unlikely... that we could write such a TM, program, or grammar in the
> absence of detailed information about how the biological system really
> functions (i.e., about its strong generative capacity, if you will). Of
> course, those who accept the Chinese Room argument are assuming ex
> hypothesi that such a TM etc. exists, since otherwise the argument is
> pointless, so this should not be controversial position.

Again, Searle accepts this position, and calls it "Weak AI": The idea that computer-modeling may be useful in helping us figure out how the mind works, though it would probably be more fruitful to study the brain directly. I happen to reject this. I think trying to pass the TTT is constraint enough. We already know what human TTT performance capacity is, roughly speaking; now we need to model it. Brain data are only useful if they suggest ways of generating performance, which they rarely if ever do. Rather, it's functional neuroscience that looks to behavioral and cognitive science for guidance as to what functions to look for in the brain. In my view, by the time the TTT is passed, all that will remain for the TTTT will be inessential implementational detail. (The premise, arguendo, of the Chinese Room Argument, by the way, is only that the TT could be passed by symbol manipulation alone. And Searle's demonstration is that, if it can, it will be passed mindlessly.)

> See what modifications in... structures produce what modifications in
> I/O behavior, ...the historical origin of the structures, etc. Thus,
> for the biological systems, we compare different people, different
> species, look at the effects of injuries and pathologies, etc., and we
> can try to replicate the relevant effects with artificial systems.

Fine strategy in principle, but so far quite unproductive in practice. Why? Because peeking and poking at brain structures does not seem to reveal function as it did in the case of livers and hearts, which wear their performance capacities on their sleeves, so to speak. The brain's performance capacity is the TTT, and I think its functional basis cannot be read off of its anatomy, physiology and pharmacology. A successful performance model will be needed FIRST, to know what to look for in the anatomy, physiology and pharmacology.

> Having said all this, I would just like to stress that the traditional
> AI rhetoric is not in fact consistent with what the theory of
> computation teaches us, that from what I have said it follows that it
> might be true (as a factual matter) that only say a biological system
> could be conscious but also that this cannot be shown by a purely

> formal argument (like the Chinese Room one), and that (although it is
> highly unlikely that we could ever write a program which could pass the
> Turing Test by focusing just on getting the correct I/O behavior
> without understanding the structural principles on which human beings
> are put together) it is nevertheless true that if such programs existed
> in any reasonable domain (say translation from one NL to another), they
> would be tremendously useful and would be considered quite correctly to
> be a vindication of the years of efforts that have been invested in AI.

This will all be music to Searle's ears. However, he does show by a formal argument (doesn't he?) that, at least in his OWN hands, pure symbol manipulation does not generate understanding. If it is a "fact" that it nevertheless does so when performed by a VAX, I'd like to know the reason why (and how you ascertain it)! Of course one can't show by a formal argument that a candidate understands (or is conscious)! The only way to do so is to BE the candidate, which is what Searle in a sense does in the Chinese Room, and the answer from there is: No! And, as I've mentioned, a vindication of Weak AI the is not necessarily a vindication of the study of brain structure as the proper road to TTT success.

> The likelihood of attaining such a program by hacking rather than doing
> empirical science of the relevant sort is of the same order of
> magnitude as that of getting such a program by typing at random. In
> linguistics, I think this idea is reasonably clear: that you cannot
> hope to write any kind of a grammar without some notion of the
> underlying structural principles, but of course linguistics is not AI.

This is a highly misleading analogy. What you should be asking about is the likelihood of attaining the "structural" principles of a grammar by studying the brain! And cognitive and behavioral theorists are no more perspicuously characterized as "hackers" than linguistic theorists are: The proper data domain for linguistic theory is our linguistic performance capacity; the proper data domain for psychological theory is our TTT performance capacity (which happens to subsume the former, as Chomsky himself pointed out).

> (For Stevan: I don't see the force of your argument about transduction,
> since every physically realized computational system (as opposed to a
> program or TM or grammar written on a piece of paper) has transduction
> of a rather simple sort and behaves as it does in virtue of the laws of
> physics (i.e. causally). But I still have not gotten your paper on this
> subject. I am just judging by the remarks you make on the subject in
> the e-mail to others that you have so kindly shared with me).

You're talking about the wrong kind of transduction. Of course transduction is involved in implementing an inert formal symbol system, written on paper, as a dynamic physical symbol system (a Turing Machine or digital computer): but transduction is not what is being implemented, symbol manipulation is! Here is the Transducer Counterargument from "Minds, Machines and Searle":

"Transduction, like flying, is necessarily nonsymbolic, drawing on analog and analog-to-digital functions that can only be simulated, but not implemented, symbolically. Searle's argument hence fails logically for the robot version of the Turing Test (the TTT), for in simulating it he would either have to USE its transducers and effectors (in which case he would not be simulating all of its functions) or he would have to BE its transducers and effectors, in which case he would indeed be duplicating their causal powers (of seeing and doing)."

> Stevan, I just reread your earlier message and realized I had not read it
> carefully enough. You say:
>
> "Strong equivalence vs weak equivalence is only relevant here if it
> can distinguish a parallel vs. a serial implementation in some way that
> might be relevant to success vs. failure in implementing mental
> properties. Two algorithms for doing a factorial may be weakly
> equivalent but not strongly equivalent, but it is unlikely that this
> would have much to do with mental capacity; and ex hypothesi, weakly
> equivalent algorithms must have identical behavioral capacities."
>
> Of course, two weakly equivalent algorithms have identical behavioral
> capacities. So do the Searle in the Chinese Room and a native speaker
> of Chinese. Now, many critics of AI and computational approaches to
> cognition, incl. I thought both Searle and you, seem to assume that
> FROM THE COMPUTATIONAL POINT OF VIEW the Searle and the native speaker
> must be equivalent, and hence the whole argument is that the
> computational point of view is mistaken. Now, without committing myself
> to accepting the Chinese Room argument (at least in its conventional
> form), I am at pains to note that FROM THE COMPUTATIONAL POINT OF VIEW
> the Searle in the Room and a native speaker need not be
> indistinguishable, since they may be and indeed probably are strongly
> non-equivalent. Hence, it is possible (indeed I think it is the case)
> that the relevant differences between said Searle and said native
> speaker are differences which the other Searle (the real one)
> characterizes in terms of intentionality and biology and you do in
> terms of transduction (if I understand aright) and which a theory of
> computation buff could characterize in terms of strong generative
> capacity or the like.

I’m afraid you’ve got it mixed up: The comparison is not between (any) Searle and a native Chinese speaker (who are no doubt strongly nonequivalent!) but between Searle following symbol manipulation rules and a VAX following exactly the same rules. They ARE strongly equivalent, and if one of them does not understand, there is no reason to believe the other one does. Besides, whether or not you commit yourself to accepting the Chinese Room Argument, you seem to be accepting all of its conclusions (even those I reject)!

Stevan Harnad

To: djoslin@bbn.com POBox 1592, Cambridge MA 02238

David, you wrote:

> The symbol grounding problem reminds me of several things by Hilary
> Putnam. (The first part of _Reason,_Truth_and_History_, for example.
> On p. 51: "How do the thinker's symbols ... get into a unique
> correspondence with objects and sets of objects out there.") Were you
> influenced by Putnam's work?

I wasn't directly influenced by Putnam, but I've noticed some parallels and connections. See the last chapter in "Categorical Perception: The Groundwork of Cognition" (Cambr. Univ. Pr. 1987) where I explicitly discuss the connection between Putnam's word/world problem and my symbol grounding problem. (Dan may have a reprint lying around. If not, let me know and I'll send you one.)

> You say (p. 15) "Only implemented mechanisms that can pass the TTT --
> i.e., respond to all of our inputs indistinguishably from the way we do --
> can understand." I understand how this is a sufficient condition, but
> I don't see why it is a necessary condition that the mechanism be able
> to respond to *all* our inputs. A person born blind or deaf or both
> doesn't share *all* our inputs, but we certainly wouldn't question
> whether their mental symbols are grounded. So if a computer passes the
> TT, and has learned much of what it knows by using a mobile video
> camera (but cannot examine things by touch, perhaps), wouldn't that
> solve the symbol grounding problem?

Both these questions have come up before. The TTT is supposed to cover our characteristic robotic powers. Handicapped people are special cases; of course they're grounded too, as long as they have any sensory equipment at all. But our sensory equipment isn't all on the outside. And what's on the inside isn't just a symbol cruncher. So the deaf still have a lot of nonsymbolic (potentially) auditory internal equipment, and who knows how they're using it. Forget about the fancy pathologies until you get a handle on the normal case, which consists of what MOST of us can do.

As to a symbol curncher with just one optical module: I'm betting that won't be enough to ground anything, and that the solution will be a hybrid system in which all functions are suffused with transduction and there is no autonomous symbol cruncher at all. But try it. See if you can get TT level performance with just a mobile visual module. I'd accept that much robotic power as being in the same ballpark as the TTT...

> Suppose, to take another approach, that we have a program that passes
> the TT but has no transducers at all. But suppose that we've written
> the program so that when it talks about apples it can form a "mental
> image" of an apple, and one that we would not be able to distinguish
> from a digitized image of a real apple. This mental image did not
> come from a digitized image of an apple, though, but is entirely
> software-generated. If you ask the program (during the TT) to imagine
> a New England scene, it might form a mental image of a (non-existent)
> church with a steeple, surrounded by snow and trees, etc, and then
> proceed to describe the scene it has constructed. Perhaps it is even

> constructed to "remember" having been to such a place during its
> (non-existent) childhood. So far, though, this approach runs up
> against the symbol grounding problem as much as any other purely
> symbolic system.

It sure does. Because you really have no "images" or "memories" in there; just symbol crunching that we can INTERPRET as images and memories.

> But now, the day after it passes the TT, we connect it to sensors. We
> take it around New England, letting it observe the scenery. It may
> even experience deja-vu when it gets out to Southborough, for example,
> and think that this is the scene it was remembering in yesterday's
> Turing Test. Is this robot "understanding" today, but was not
> understanding yesterday? If we now disconnect its sensors, does it
> lose its ability to understand?

NOW you're dreaming the impossible dream of the symbolic functionalists, the one my papers are devoted to rousing them from: They think that the real cognitive work is just symbol crunching, and that "grounding" is just a simple matter of hooking up the symbol cruncher, via peripherals, to the world, "in the right way." Well Putnam noticed that that "in the right way" business may not be as simple as it seems. And my whole point is that the kind of bottom-up grounding scheme that I'm proposing can't just be a hook-up between two independent modules! It's bottom-up all the way. There's nothing to separate and "disconnect," in the way you unhook a symbol cruncher from its transducers and effectors. That's what grounding is all about.

> Similarly, suppose we take a human brain and sever all of its sensory
> inputs -- the "Brain in a Vat." Does the mind there now start to have
> symbol grounding problems? I haven't been following all of the
> discussion in the e-mail (and I missed all but the very tail end of the
> comp.ai discussion) so perhaps these issues have been covered before...
> I'm taking Dennett's "Philosophy of Mind" class at Tufts this semester,
> and thinking about doing my final paper on the symbol grounding
> problem, comparing your response with Putnam's and Dennett's.

Yes, these issues have been discussed before. And the brain-in-a-vat example is discussed explicitly in "Minds, Machines and Searle." If you're going to do a final paper on the symbol grounding problem I think you should read not only "The Symbol Grounding Problem," but the two other papers I mention here, both cited therein.

The short answer about the brain in the vat is that it's a fantasy to think that it's just a symbol cruncher. Vast portions of the brain, apart from the sensory surfaces themselves, are simply analog (nonsymbolic) re-presentations of the sensory surfaces and various analog transformations and reductions of them. If you cut off the sensory surfaces you have not cut out the sensory system. And if you keep cutting, so as to remove every piece of tissue that's either sensory or motor, you'll make a big mess, you'll have very little brain left, it'll look more like egg drop soup than a brain in a vat -- and you STILL will not have isolated a symbol cruncher!

> I noticed that you are giving two talks at Swarthmore. I wish I could
> go, but it’s a bit far. Any plans to give the same talks in or closer
> to Boston?

I’ll be giving two talks at MIT on Jan 25th.

> I’m currently looking around at PhD programs in Cognitive Science, or
> any program that combines AI and philosophy. Can you tell me something
> about the "Behavioral and Brain Sciences" department, or other Cog Sci-
> related work at Princeton?

BBS is a journal I edit. The department here is psychology, and there is a cognitive science program. The ones who can give you information are: Brian Reiser bjr@clarity.princeton.edu or Gil Harman ghh@clarity.princeton.edu

Stevan Harnad

-----

miken@ai.mit.edu (Michael N. Nitabach) wrote:

> Subject: Cognitive Science Methodology
>
> a peripherally deaf person still has a great deal of neural tissue
> which was originally devoted to auditory signal processing. It is an
> empirical question, however, to what extent this tissue can play any
> role in mental function, given the absence of peripheral auditory
> input. My intuition is to think that this will be highly dependent on
> when, and how, the individual became deaf.

Look, I agree with all this; I myself am a neuropsychologist (I used to work on lateralization). The neural tissue-issue is just to stave off the hackers’ fantasy that once you’ve removed the sensory surfaces, there’s nothing left but a symbol cruncher. Of course there’s plasticity, and early learning effects, etc. It’s just not related to the issue at hand.

> even though the brain in the vat has a whole lot of neural tissue
> normally devoted to sensory signal processing, it is not clear that
> this tissue would be active at all, or do anything, without peripheral
> sensory input. What also seems questionable is whether this tissue, if
> it never developed with appropriate sensory input, would perform any
> useful function if sensory input was later established.

These are all empirical questions about plasticity, autonomy and modality-specificity: Interesting, but not relevant to the issue under discussion, because symbolists always think the brain in the vat is just a symbol cruncher cut off from its peripherals, and I’m pointing out that most of the brain, whether cut off or not, is just peripherals too. I’m not committed to anything more here.

>> "Forget about the fancy pathologies until you get a handle on the >> normal case, which consists of what MOST of us can do." >
> I wholeheartedly and completely disagree with this statement. It is

> just a consideration of the pathological cases which provides insight
> into (1) the assumptions of our theories, and (2) the constraints which
> reality imposes on our theories [e.g., neuropsychology, physics]...
> I suppose that it is possible to consider your statements as a
> pedagogical suggestion. i.e. "In developing your ideas, understand
> their application to typical situations before applying them to the
> unusual." As such, I can't really quibble. However, it is crucial to
> eventually apply one's ideas to the "pathological" cases...

Actually, it's not just a pedagical suggestion. I'm struck by the fact that neuroscience has produced so few ideas (if any) about the functional mechanisms underlying our behavioral/cognitive (i.e., TTT) capacities. Of course if any ideas about how to generate performance did come from neuroscience I'd happily listen to and learn and apply them. They're just not coming. The arrow seems to be going the other way: Functional modelers are giving neuroscientists an idea of what to look for. I think this is partly because of the usual division of labor, as in physics, between theory and data gathering, neuroscience consisting largely of the latter. But it may also be that with the kind of reverse engineering involved in trying to understand brain power, Claude Bernard's strategy of looking at the pathologies of biological organs in order to understand their normal function just doesn't work, because behavioral capacity -- especially TTT scale -- is such a special kind of "function."

In practice, it has not been lost on some that, behaviorally and cognitively speaking, neuropsychology in particular has been the science of studying (and localizing) functional deficits, not functions. And on the other side, toy models of performance have tended to mask their own "deficits" -- i.e., their pathetically puny scale, compared to the TTT -- by playing on our credibility with little brain-like frills, like "neural" nets, "graceful degradation," etc. (Aping some of the fine-tuning parameters of performance, such as reaction times and error patterns while offering infinitesmal fragments of competence, falls in the same category of cover-up for performance deficits.)

One of the things I learned from laterality work was how easily you can inspire almost mystical reverence with a trivial result (such as faster reaction times in the left visual field to low spatial frequency stimuli) merely by linking it to "the brain" (and hemisphere myths in particular). The same RT difference under, say, different verbal instruction conditions would instead elicit the vast "ho hum!" response it deserves. Well, the same thing happens when you take a trivial performance model and and find in it (or give it) some trivial brain-like property.

That's why, for now, I think performance modeling should just focus on normal performance, and not worry about gussying it up with pathology. I'm not saying we shouldn't take the insights offered by pathology, if ever it offers any insights. But meanwhile, the only performance constraint worth aspiring to is to "scale up" to the TTT.

Stevan Harnad

[To: recipients of the Symbol-Grounding List. Please let me know if you wish your name removed from this list -- or someone else's name added. -- SH]

From: Stevan Harnad

TURING EQUIVALENCE VS THE TURING TEST

Alexis Manaster-Ramer amr@ibm.com wrote:

> Searle and other critics of AI seem to want to explain, in part, why AI
> has not succeeded, but in fact their arguments (such as the Chinese
> Room argument) presuppose that AI HAS succeeded (in its own terms)
> precisely by providing such programs. If, as Harnad says and as I argue
> also, such programs are not possible, then the line of argumentation
> chosen by Searle is not the right one. For, in fact, Searle here makes
> essentially the same assumptions about what is possible as Minsky, say.

No, both Searle and I consider the assumption (that the TT could be passed by symbol crunching alone) "arguendo." I actually thinks it's counterfactual, and I expect Searle does too. The logic is this: Part of the analysis of the consequences of the assumption's holding true, were it not counterfactual, casts considerable light on WHY it may be counterfactual. Since, though it may be counterfactual, it is not clear that it is provably impossible or incoherent, an argument based on supposing it to be possible is not incoherent either, and has even proven very informative (regarding the ungroundedness of pure symbol manipulation, and the weaknesses of the arguments, such as the "Systems Reply," adduced in its defense). The counterfactuality is discussed explicitly at the end of "Minds, Machines and Searle."

> to simulate a paralyzed speaker of a language but not an able-bodied
> one cannot count as a successful model of ALL possible speakers of the
> language.

And that's just one of the many reasons for insisting on the TTT (Total Turing Test: all our capacity for sensorimotor interactions with the things in the world, including linguistic interactions) rather than just the standard TT (linguistic interactions):

(1) There's no reason to suppose that a partial model, with only the capacity for a toy fragment of our performance, would have a mind.

(2) Scaling up to the TTT reduces the degrees of freedom and may converge on a unique, or near unique (within the limits of normal empirical underdetermination), family of functions (the "convergence" argument).

(3) There are reasons (arising from the Symbol Grounding Problem) to believe that the capacity to pass the TT must be grounded in the capacity to pass the TTT.

(4) The TTT, not the TT, is the basis of our everyday, informal solution to the other-minds problem.

(5) Apart from the TTTT (i.e., the TTT expanded to include brain "performance," which will, I think, supply only fine-tuning or implementation-specific information once the TTT is passed) there is no further empirical basis for designing or testing a theory of mental function, because of the other-minds problem.

> I think there are two different issues of grounding, one of which has
> to do with the fact that people are expected to have some interaction
> with the world, the other with the question of whether a computer or
> its ilk can have cognitive abilities (and disabilities) like those we
> think we have. The former can be possessed (in principle) by a robot
> without in the least affecting the latter question, and so I think that
> a Searlean argument could still be made. However, again, I think that
> it is even less likely that we can build a robot that behaves like a
> human being in all respects but does so without replicating in any way
> the basic [i.e., mental] way in which people do things.

No, Searle's argument won't work against a robot that passes the TTT, because an implemented robot necessarily draws on nonsymbolic functions such as transduction, which, unlike mere symbol manipulation, Searle himself can neither perform in the Chinese Room, nor leave out, on pain of the "systems reply."

I repeat: You can't do better than a grounded robot that passes the TTT (apart from fine-tuning it with the TTTT to include its brain's "performance"). And there are exactly three reasons to strive for symbol grounding: (1) It would provide the link between meaningless symbols and the objects and states of affairs they represent; (2) it would provide the wherewithal to pass the nonverbal component of the TTT, making performance completely indistinguishable from our own; and (3) it would provide natural immunity against Searle's Argument.

Your last point is discussed as the "Convergence Argument" in "Minds, Machines and Searle."

> Finally, I would just like to ask you to imagine a program or robot
> which passes the TT or the TTT, and since its I/O behavior is the same
> as Searle's, it proceeds to argue (in print even) that it possesses no
> understanding, no feelings, etc. Would you take its word for it? Should you?

I would personally be inclined NOT to take its word for it, but this is a tricky question, perched on the brink of paradox. To get the flavor of this, consider that, since it passes the TT, if I ask whether it understands what I'm saying, it must say "yes." (Any candidate that merely kept saying it didn't understand me would quickly fail the test.) But at the same time that it affirms that, like you and me, it understands me, it must also deny it, almost in the same breath. ("Yes, I understand, but not REALLY.") Similarly with seeing a sunset: ("Yes, I see it, but not REALLY; in reality I'm just going through the motions.") And with any mental state: ("I think, want, believe, imagine, hope, feel... but not REALLY...") I suppose we could string out a sci-fi scenario along these lines for a while, but it seems even more far-fetched and counterfactual than the assumption that a symbol cruncher alone could pass the TT.

It is tempting to infer -- from the fact that it is so hard to believe or even conceive that there could be a coherent and credible way to speak and act exactly as we do and yet deny all vestiges of subjective life -- that this may have something to do with the functional utility of consciousness itself. Yet, on our own hypothesis (i.e., complete TTT [or even TTTT] equivalence) there seems to be no functional room for real consciousness to confer any causal advantage WHATSOEVER over mere as-if consciousness! Hence, apart from the pleonasm, I don't really see why there should be that vast difference between us and a zombie that is just going through the motions (though I of course know there IS that vast difference). Hence perhaps my inclination not to believe the robot's denial that it has a mind may just be a consequence of the natural intuitive compellingness of the TTT for all of us. What’s sure is that we are up against the other-minds problem either way.

I, by the way, am a methodological epiphenomenalist: I think subjectivity is a "spandrel," with no independent causal power or functional utility, a fellow-traveller that happens to piggy-back on any mechanism that has TTT (or perhaps TTTT) power.

> I emphatically deny that the theory of computation makes any
> distinction between algorithms and implementations.

This is a distinction whose defense I will have to leave to conventional computationalists and functionalists. It seems to me that the software/hardware distinction and the notion of multiple realizability are quite solid. In any case, that is not the turf I am either defending or attacking. If the software/hardware distinction is untenable than both Searle and Strong AI are wrong, and every system is unique and incommensurable.

> we do not know, given a machine, even approximately what we are to take
> its "input" and "output" to be, unless, of course, someone tells us.

I just can’t understand why you say this, because for every real organism or machine and all real inputs and outputs I can think of, there’s no problem at all. The I/O of a rat in a Skinner box is readily enough specified. Likewise for a pair of speakers exchanging words, or for a person manipulating the objects in the real world. Moving to artifacts: The old IBM mainframes used to get punched cards as inputs and wrote paper as outputs. This VAX takes keystrokes in and gives CRT patterns out. Searle gets Chinese symbol strings in and sends Chinese symbol strings out. And robots operate on objects, as we do.

So there’s no problem with WHAT the I/O is in any of these cases. Now if you are referring to the INTERPRETATION of the I/O -- what it means, or what have you -- that’s another matter: It’s the symbol grounding problem, in fact, but that’s exactly what’s at issue here! Until it’s solved, the inputs and outputs can certainly be specified, but only as meaningless squiggles and squoggles -- tokens in a symbol system whose interpretation comes from US rather than being grounded in the system itself. (Or perhaps you mean the problem of anticipating and simulating I/O contingencies in a computer model of a robot: But that too is part of the symbol grounding problem.)

I suggest that you distance yourself from unnecessarily paradoxical positions such as our not even knowing what I/O is, because they lead only to paralysis.

> Now an implementation of an algorithm is not a formal concept
> at all, but a practical one, and whether something is or is
> not an implementation of a given formalism is not always clear.
> This is precisely why there is no one notion of equivalence that is
> relevant. For example, if determinism is at issue, then an
> implementation that does not preserve it is going to be useless.
> Also, in this context it is often noted that a nondeterministic
> machine (in the technical sense, not in the sense of something
> that behaves randomly as in the cases of quantum mechanics)
> can have no physical realization.

Relevant to what? You seem to have lost sight of the question at issue here: The question at issue is whether or not a pure symbol cruncher that passes the standard TT (symbols in, symbols out) can understand the symbols, and in particular, whether or not Searle's Argument has shown that it cannot.

One can certainly formalize the CONCEPT of implementing an algorithm (that's exactly what the turing machine concept does); what is not formal is the implementation itself: That is always a dynamic physical system (hence also "practical"). It is still true, however, that there can be two DIFFERENT implementations of the SAME algorithm, and that this is usually called "strong equivalence" (in contrast to "weak equivalence," in which the I/O is the same but the algorithm differs).

One could define a third, still "stronger" form of equivalence, I suppose, namely, "implementational equivalence," based on considerations about the physical structure of the dynamical system implementing the symbol crunching, but it's not clear how these are relevant to this discussion. For it is already true a priori that if having a mind requires IMPLEMENTATIONAL equivalence with a biological brain exactly like ours, then only creatures with biological brains exactly like ours can have minds. That, however, sounds to me more like a premise than a conclusion, and hence incapable of settling the Chinese Room Problem. (Note that even Searle does not demand implementational equivalence -- only equivalence in whatever of the brain's "causal powers" are necessary for having a mind; this too is a premise, but a self-evident one, it seems to me.)

Finally, for unrealizable machines (e.g. unrealizable nondeterministic ones), the implementational-equivalence question seems moot altogether.

> all we ever have is two (or more) systems which, while distinct, are in
> some sense equivalent. And the problem is: Are they equivalent in the
> relevant sense?

All this talk about systems in general and equivalence in general is too vague. What's at issue here is pure symbol systems only, and whether they can understand. The premise is that implementing the right set of symbol manipulation rules amounts to understanding; Searle shows that's false. The second order question of implementing the right rules in the "right way" seems too vague, and certainly, as I said, cannot settle the Searle issue.

I also think you're confusing Turing equivalence with passing the Turing Test, as in the following:

> even the abstract models can be interpreted as computing different
> functions (i.e. passing different Turing Tests!!!) depending on our
> definitions.

Turing equivalence refers to computing the same function (weak) or computing the same function with the same algorithm (strong). In both cases, the equivalence is with some formalizable turing machine state and turing machine table. The Turing Test, on the other hand, is an informal test of whether a human being can tell apart a human being from a machine by their I/O alone. There is only ONE TTT (though it may have homologues and precursors for other biological species).

The problem of the "definition" of the meaning of the I/O, etc., is again the Symbol Grounding Problem.

> In the case of a physical system, I suppose, you could provide
> some observational definition of what we view its I/O as being,
> but then it would not be true that the same program when run with
> two different output devices, computes the same function. You have
> to take the human observer as part of the system which performs a
> conversion from the actual output to some standard representation.
> Whenever we talk about implementation, we assume some such mental
> conversion. But strictly speaking, we must consider every implementation
> to be a quite different machine from the algorithm which we find it
> convenient to say that it implements. This means that the crucial,
> and mostly open, question is exactly what sort of equivalence we
> want for various purposes.

Again, this seems to me to be needlessly vague and mysterious. Even allowing for the symbol grounding problem, there are certainly ways of ensuring that the I/O of two different systems is not only equivalent, but identical. This is perfectly observer-independent. The observer-dependence only comes in when we INTERPRET the I/O or the internal states of the system, if it's a symbol cruncher. I don't think your worries about "sorts of equivalence" is picking out a coherent problem here.

> a system that behaves (assuming we have a definition of this) like
> a human being [and] is physically realized in THIS universe, and is the
> same size as a human being [may] have to be made of water.

As I've said, there are no "definitional" problems for the performance criteria of the TT or the TTT. And the water argument is just a repetition of the possibility that some (unknown) constraint may make only the exact biological brain eligible. Maybe so, maybe not. Mind modeling will be testing empirically what functions will and won't give a system the capacity to pass the TTT.

> where I disagree with Harnad [is that it] is not true that you need to
> actually interact with the physical world, you need only be able to
> talk about how you would interact if you could.

Yes, but in "Minds, Machines and Searle," and "The Symbol Grounding Problem," I give the reasons why having the functional wherewithal for being "able to talk about how you would interact [with the physical world] if you could" may well depend completely on having the functional wherewithal "you need to actually interact with the physical world," and probably to have used it some too. That's the grounding proposal.

> simulating the world for the intelligent machine to interact with would
> probably be an even worse task than that of building the machine
> itself.

Though simulations can anticipate a lot, I think one can as little avoid developing and testing a TTT candidate on real-world I/O as one can avoid giving a simulated airplane some real flight experience.

> why, after all the centuries of working natural science and
> mathematical models thereof, is there still no formalism (implemented
> or not) which passes the "Turing Test" with respect to the physical
> universe we live in (or even a small chunk of it)...

I don't see the point here. Physics is not a part of engineering (though engineering is a part of physics, and a rather successful part). Physicists don't actually build universes, though they do try to find their "algorithms" in the sense of invariances or natural laws (sometimes even simulating them). And I would say that Newtonian mechanics, quantum mechanics, and special and general relativity have done a pretty good job in second-guessing a goodly portion of the universe's "behavior"...

Mind-modeling is the only field that has a Turing Test. And, as I pointed out in "The Symbol Grounding Problem," it will always have a second order of uncertainty, over and above the usual empirical underdetermination of all the sciences. For whereas when the physicist is asked: "Yes, your theory fits all the data, but how do you know it's really TRUE of the world? How do you know the universe really hews to THOSE laws rather than some other ones that likewise fit the data? Is this not a mere look-alike of the universe?", it is quite reasonable to reply that that difference, such as it is, can only be fought out in the arena of actual competing theories. Beyond that, a false look-alike is only a conceptual possibility that cannot make a palpable difference to any of us. That's ordinary underdetermination.

But when a mind-modeler is asked the same questions: "Yes, your model passes the TTT [or even the TTTT, if you wish], but how do we know that it's TRUE of us? How do we know it really has a mind, rather than being a mere look-alike of something with a mind?", he must concede that this is worse than ordinary underdetermination. For there is at least one being to which having a real mind, versus being a mere lookalike that is going through the mindless motions DOES make a difference, and that is the model itself. And each of us, as we contend with the other-minds problem. Searle's Chinese Room Argument taps this special feature of the mind/body problem directly.

Stevan Harnad

THE SYMBOLIC HALL OF MIRRORS, AND HOW TO AVOID IT

The following text is from:

David Chalmers Center for Research on Concepts and Cognition Indiana University.

It will be accompanied by my annotations in square brackets - [...].

Stevan Harnad

From: Subject: Deconstructing the Chinese Room.

Circulate and reply publically if you like. If you do so, I'd prefer that you send my comments as is, followed by another message with your reply (and whatever excerpts you like). I find that it is extremely hard to get the gist of an argument when it is presented solely in excerpted form (those ">"s make a difference, too). Alternatively, if you don't want to send two messages, present mine as is with your comments enclosed in brackets. Well, you can do what you like, but that's my preference.

- [My comments will be in square brackets. SH]

This represents my considered position on the Chinese Room, and on your accounts of it and variations. I have tried emphatically to *not* just repeat the same old arguments, but to get at the real meat behind the arguments, and in particular at the intuitions which support them. I think I have presented four rather difficult challenges for you to answer, if you are going to continue to defend your present position.

I think I understand your argument perfectly, but I still don't buy it. Here are a few points of issue: quite separate points, as indicated by the numbering. The four points relate to:

(1) The causal intuition; (2) The semantic intuition; (3) Computers and the Turing Test; (4) The Total Turing Test.

(1) I am less interested in the Chinese Room argument itself than in the intuitions behind it, that lead some people to accept it. There are two of these, it seems. The first, which you clearly share, I call the "causal intuition." When you dismiss the Systems Reply (which is, of course, the Reply that I would give), you give reasons like: "But the human in the Room is doing *all the work*. If they don't understand, then nobody does."

Now, this argument is implausible to me and others, as has often been stated. Clearly, most of the complexity of the system lies not in the human but in the external apparatus (paper, symbols and rules). In fact, the job of the human could be done by a very simple automaton (a CPU, if you like). The fact that the human is "the kind of being that *might* understand something" is just a distraction; we shouldn't expect the understanding in the system to supervene solely on the human. But anyway: clearly you've heard this argument before, and don't buy it. Why not? Why can't you accept that the system as a whole understands?

- [SH: Because, (1) as he's stated repeatedly, Searle could in principle - memorize all that stuff, and then there would be NO EXTERNAL APPARATUS - at all! Besides, (2) it was never plausible in the first place that - if Searle alone couldn't understand then /Searle + paper-with-symbols/ - could understand. "System-as-a-whole" is just hand-waving at the - ghost that hackers** have become all too accustomed to freely - projecting onto their machines, first just as a mentalistic - short-hand, by way of analogy with mnemonic variable names in - higher-level programming languages, but eventually as a habitual and - uncritical hypostasization that some have even come to regard as an - "axiom." - ---

- (**And not just hackers, apparently: even sophisticated philosophers - like Dan Dennett, whose celebrated epiphany when he remarked that the - chess-playing computer program "thinks it should get its queen out - early" has led him to believe that there is NO DIFFERENCE between real - thinking and interpretability-as-if-thinking: It's all just a matter of - explanatory convenience in explaining behavior. (Perhaps, then, there is - no difference between real flying and the "interpretability-as-flying" - of symbolic simulations of flying either, and that's all just a matter - of explanatory convenience too?) The problem is that this explanatory - convenience only matters and means anything to creatures that really - think. And Searle is here to remind us that the difference between - understanding Chinese and interpretability-as-if-understanding-Chinese - is very real indeed, because he clearly has only one and not the other. - Besides, semantic intrepretability is already one of the DEFINING - PROPERTIES of formal symbol systems, and we surely can't help ourselves - to a mind by definition, can we?) SH.]

Before you try to answer that, I'll answer for you. The only reasonable answer, it seems to me -- the only *principled distinction* between the human and the rest of the system -- is that in a sense, the human is *causally active* and the rest of the system is *causally passive*. To give background: the great appeal of the computational model of mind for many of us lies not in computation per se, but in the fact that computation allows us to model any *abstract causal structure* we like. It is this fact that allows us to step from abstract functionalism to computationalism; the fundamental belief is that what is important to the mind is its causal structure, and the computer allows us to model this. In the Chinese Room, such a complex causal structure is also present, but you do not even consider imputing mentality to this structure. Your sole reason comes down to: the human is doing all the work.

- [SH: As you see above, that's not how I would answer. The "abstract - causal structure" of an airplane can also be modeled by a symbol - cruncher, yet it doesn't fly. (Right?) So why should the "abstract - causal structure" of a mind (if you managed to simulate all of it) - think? The only thing an abstract causal structure can do (if - implemented as the rule table for a physical symbol system like a - turing machine or digital computer) is to guide the crunching of the - meaningless (but semantically interpretable) symbols in which the - abstract structure is encoded. This "ungroundedness" of symbol - systems, as I've called it, is what Searle successfully exploits in - the Chinese Room Argument. SH]

Now, there's no question that in the CR, it is true to say that "symbol X equals such-and-such" *causes* symbol Y to be accessed, which causes... The causal structure is just as complex. The *only* distinction, perhaps, between the causation here and that in a brain, is that this causation has to be *mediated by an active agent*, namely the human. Whereas brain-style causation is "direct". When you say "the human is doing all the work" you mean precisely "the human is supplying the active causal wherewithal". Therefore: you must hold that there is a *principled distinction* between active and passive (mediated) causation, and that this distinction is sufficient for vast metaphysical distinctions: in particular, that this distinction is sufficient for you to dismiss out of hand the idea that such a complex system might understand. Whether you hold this intuition explicitly or not, I hold that it is the only route open to you in justifying your dismissal of the Systems Reply.

- [SH: The brain is a red herring here; no one knows what it does. - And the issue is not "active vs passive causation." What's at issue - is Searle and a pure symbol cruncher, and the punchline is, whatever - the brain is doing, it can't be just symbol-crunching: - - Look, we both agree, presumably, that while the code is sitting - inertly on the disk and the CPU is idling, neither a symbolic - simulation of flying nor a symbolic simulation of thinking is causing - anything to speak of, right? Now lets run each of them on a symbol - cruncher: On the face of it, there's now just symbol crunching going - on in both cases. That's clearly still not flying. Why should it be - thinking? - - In "Minds, Machines and Searle" I distinguish two "kinds" of causality, - one the usual physical causality, and the other mere formal - "schmausality," of the kind you find in interpreting an

"If A then B" - statement. There's clearly physical causality going on when a - computer is executing a program, but it's the WRONG KIND of - causality, both in the case of flying and thinking. What's happening - in both cases is that symbols are being manipulated on the basis of - syntactic rules operating on their shapes. Now the whole to-do is - indeed interpretable-as-if flying or thinking were going on, because - systematic semantic interpretability, as I've already noted, is one - of the defining features of a symbol system. So what? Even if, in the - case of the flight simulation, all the abstract principles of flying - have been captured symbolically, it's still just symbol crunching. - Why should it be any different in the case of thinking? - - Pure symbol manipulation lacks the requisite causal power to implement - the function in question in both cases; it can only simulate it. And, - as likewise noted in "Minds, Machines and Searle," the standard Turing - Test (TT) (only symbols in, symbols out) is SYSTEMATICALLY AMBIGUOUS - about the causal difference between real mental processes and mere - "interpretable-as-if-mental" processes. Only the Total Turing Test (TTT) - (full sensorimotor interaction in the world) can capture the causal - difference. After all, a simulated airplane could pass the symbols-only - "TT" too; the real test is flight, in the real world. Why should we ask - less of minds? SH.]

Now, clearly if you're going to hold this [active/passive] intuition, you had certainly back it up. While I can't say that it is analytically disprovable, it's certainly not intuitive for me, or others. In fact, philosophers trying to elucidate the notion of cause have generally come to the conclusion that such distinctions are impossible to make, and to uphold in a principled fashion. Causation is causation. So, this is my first request: justify this intuition.

- [SH: As Aristotle said: There are formal causes and efficient causes. - You can't implement a mind with just formal causes that are merely - INTERPRETABLE as being more than that. And, as I said, it's in any - case not the active/passive distinction I'm invoking, particularly, - but the distinction between mere symbol manipulation, be it ever so - active, and other, nonsymbolic processes.]

Incidentally Searle, strangely enough, *rejected* this [active/passive] intuition implicitly when he rejected the Haugeland's Demon argument (about the demon which ran around doing the causal work of broken neurons). He claimed that such a system would still understand. This was strange, because it meant that he rejected the most plausible defence of the "but the human is doing all the work" argument.

- [SH: I don't know what neurons deep in the brain do or cause, - but I've given reasons why it's not just symbol crunching. The - issue is not activity versus passivity. Searle and a symbol - cruncher can take over functions or parts of functions that just - amount to simple causation, such as symbol manipulation. Neither, - however, can take over ALL of an essentially nonsymbolic function - such as heating, flying or transduction. To the extent that neurons - do such essentially nonsymbolic things, neither Searle nor a symbol - cruncher could take over all of THEIR functions either (except, of - course, by BEING the kind of thing he's taking over the functions of, - or its causal equivalent, in which case he certainly would be - heating-flying-transducing). What Searle was rejecting was the - homology between brain function and symbol crunching, and their - corresponding take-over possibilities. He was also stressing that - it's not synthetic take-overs of natural functions that he's - sceptical about, but pure symbol crunching, no matter who or what - implements it.]

Incidentally 2: a parallel implementation of a connectionist network would not be vulnerable to this intuition, though its serial simulation might be. Although *any* implementation is active on the bottom level; atoms, for instance, don't make such distinctions; everything is active to them. Which is one reason why this intuition is difficult to defend.

- [SH: Passive vs. active is not the issue, symbolic vs. nonsymbolic - is. And, as noted in "Minds, Machines and Searle," parallel processing - would be as immune to Searle's Chinese Room Argument as transduction - is, but ONLY IF there is something essential to the parallelness - itself in implementing a mind. However, since it seems to be agreed - that serial simulations of parallel nets are functionally equivalent - to them (I recently queried the Connectionists email list explicitly - about this), the burden is on the essential parallelist to show - what's intrinsically special about parallelism in implementing a - mind. I have suggested to Searle a "3-room" variant of his Chinese - Room Argument (1: a parallel net passing the Chinese TT; 2: a serial - simulation of 1; 3: a Searle simulation of 2) which he will be using - in his forthcoming Scientific American article. This argument is as - decisive against nets as the original was against symbol crunching - (for the very same reasons), except if there is a functional - counterargument for essential parallelism.]

(2) That was only number one. The number two intuition is the *semantic intuition*, which I'm sure you hold. This goes, as Searle has put it: "Syntax isn't sufficient for semantics. Computer programs are syntactic. Therefore they are not sufficient for semantics, and thus for minds." Your version of this seems to be: "Computers can only manipulate symbols. But these symbols are not *grounded* (that is, they carry no intrinsic meaning). Therefore computers have no semantics."

The first thing to be said about this is that you're using the word "symbol" in two different ways here. For symbolic AI, these two uses coincide. For connectionism, by contrast, they don't. The first sense (I've said this before) uses "symbols" to represent "objects manipulated by programs." The second sense uses "symbols" to mean "something which denotes (or which claims to denote)". The second is the traditional semiotic meaning; the first is a usage more common in computer science. (Your JETAI paper, for instance, uses "symbol" in both of these senses without making a distinction.)

- [SH: There's only one sense. It's only your use of "representation" - in two ways that gives the impression of a distinction: Symbols - don't "represent" objects that are manipulated by programs. Symbols - (actually, physically tokened or implemented symbols) simply ARE - objects manipulated by programs. "Representation" hasn't come into the - picture yet, except at the metalevel where you're referring to what - our term "symbol" represents. Now let's turn to what these symbol - tokens themselves represent; that's where representation is involved - in a formal symbol system: The symbols and the symbol combinations - are SYSTEMATICALLY INTERPRETABLE as representing something (usually - objects and states of affairs, e.g., things, numbers, propositions, - events, processes, etc.). This SINGLE sense of symbol and representation - is common to symbol crunching and connectionism. What's in dispute about - connectionism (and not everyone agree with Fodor and Pylyshyn that - nets have this deficit -- e.g., Cummins thinks nets are just a special - form of symbol crunching) is whether connectionist representations - have all the combinatory sub-parts that symbolic representations can - be decomposed into while preserving systematic semantic - interpretability. By my lights, however, the semantic interpretation - is ungrounded either way -- except that in the absence of - systematicity and compositionality the semantic interpretation itself - may be untenable a priori.]

In symbolic AI, the symbols which are formally manipulated are the same as those which denote. In connectionism, the "symbols" which denote lie on a completely different level. At the level of formal computation, no denotation (semantics) is present or intended. Formal computation goes on at the level of the unit and the connection weight -- inherently non-semantic objects (except in "localist" models). The *semantics* in a connectionist network lies at a higher level -- at the level of a distributed representation, for instance. *These* "denoting symbols" are not formally manipulated at all (and are thus not "symbols" in the first sense). Instead, they are emergent from the formal layer. Such emergent representations no longer suffer from the problem of being *primitive atomic tokens*; they now have a large amount of pattern and structure within them, which in a sense *grounds* them and might allow them to have content. (Just how they might do this is an open question, as is the question of how representations in the brain are grounded in meaningless neurons.)

- [SH: All this talk of "emergence" and "higher levels" is just - fantasy. Both in the case of straight symbol crunching and in the - case of "distributed" representations, the representation is - ungrounded: The connection between the representation and the thing - represented is just in the mind of the interpreter. Don't get - carried away with the epicycles of meaning that projecting a semantic - interpretation onto something will always get you: The interpretation - must be grounded on its own, without your projection. Otherwise you're - just creating a symbolic hall of mirrors -- projecting and reflecting - meaning back and forth while obscuring the fact that it's source was - just YOU in the first place! I can say that I have here a net that - "represents," in distributed fashion, that "the cat is on the mat," - but where's the connection between that little pulsating thing and - cats, mats, cat-on-matness, and anything else? - - It won't do to say, like Humpty Dumpty, that it represents them - because I interpret it as representing them -- or even because it is - capable of SUPPORTING my systematic interpretation (which, by the - way, is only true of nets if they are indeed symbol crunchers after - all). The representation must be intrinsic, i.e., grounded. My own - proposal happens to be that if the model can actually PICK OUT, - describe and interact with (at the TTT level, not a toy level) all - the real-world objects and states of affairs that its symbols - allegedly represent, then the representation is not merely alleged - but real, and grounded. Symbol crunching alone cannot do this; it - depends on a link with the real world. And, as I've argued over and - over, the link itself is the main problem of cognitive modeling: It's - not just a matter of hooking up an autonomous symbol crunching module to - autonomous transducers. To solve the Symbol Grounding Problem, the - "connection" will have to be much more intimate and substantive than - that. The grounded robot that passes the TTT will have to be a - dedicated hybrid system, in my view, with no functionally isolable, - purely symbolic module to speak of. SH.]

This, I hold, is the force of Searle's intuition. He argues, implicitly, that *primitive atomic symbols* can carry no intrinsic denotation (the mapping from their form to their content is arbitrary). This is brought out by his repeated mention of "meaningless squiggles and squoggles." This is also brought out by your occasional mention of trying to understand Chinese with a Chinese-Chinese dictionary. Here, the (denoting) symbols are *functionally atomic*, and thus seem to carry no intrinsic meaning. Further, the only functions which are available on such symbols are *within-level* operations (such as Ch-Ch translation, or composition). This can be taken as a strong argument against symbolic AI. Atomic symbols, this intuition claims, can have no *intrinsic intentionality*. (Whether *extrinsic* intentionality exists is a difficult matter which I will not get into now.)

- ["Extrinsic" intentionality is just semantic interpretability; - everything has it, even clouds: Ask Polonius...]

The relevant point is that this is no argument against connectionism. In connectionism, symbols are not atomic objects. They have form, and substance: it is plausible that there is a natural mapping from their form to their content (remember, for intentionality you don't need *reference*, just *intension*; "pattern" might be another word for this. Ref Putnam 1975, of course). In particular, they are not vulnerable to our intuitions about syntax and semantics, for all these intuitions come from cases where the syntax and the semantics *lie on the same level* (linguistics is the obvious source of such intuitions. Words, as formal syntactic primitives, do not carry meaning.) Where the semantics inheres at a higher level, all bets are off.

- [SH: This again sounds like a flight of interpretative fancy to me, - and another symptom of succumbing to the symbolic hall of mirrors: - "form," "substance" -- what are those? Mappings don't help, if they - must be mediated by your head! (Cummins seems to think that - isomorphism by itself is sufficient for representation; maybe so. - But for mind-modeling we need intrinsic representation, not - representation mediated by someone else's mind. Hence it must be - grounded directly in the objects and states of affairs it - represents.) Who knows what "intentionality" is? Who cares? For me - it's enough that I know what it is to mean something, because I know - what it's LIKE to mean something (that's the way I know I understand - English but not Chinese). A "pattern," be it ever so interpretable, - is not enough. I need a connection between the pattern and whatever - it is a pattern OF.]

This argument applies, incidentally, just as well to a serial simulation of a connectionist network as to a parallel implementation. It works equally for other emergent models. It doesn't *prove* they have semantics, but it protects them from the intuition which is being called upon.

- [SH: I've always suspected that parallel and serial nets would sink - or swim together, but on what you've told me so far, it looks as if - they'll just sink.] (3) So far we've laid out the intuitions behind Searle's "intuition pump," and seen where they might or might not be valid. Now, I'll get a little more specific, and address your claims. First, your claim that a computational entity could not "even" pass the (linguistic) Turing Test.

This, I claim, is demonstrably false. The argument is simple, and lies in the possibility of arbitrarily good simulations of any physical system by numerical methods. Now, it's true that we don't have all the laws of physics worked out yet, so this isn't practical tomorrow. But for you to claim that it will *never* be possible would be to make an extremely strong claim, and not one I think that you want to base your argument on. Essentially, it seems plausible that nature obeys a set of (probably second-order) differential equations. If we take a system, and make a close enough approximation of its initial state, we can numerically simulate the way it follows the laws of physics over time. If our system is a human being, we can in principle digitally simulate its behaviour over time arbitrarily closely. (Arguments from "chaos" may be tempting here, but would be misguided. Chaos makes a system unpredictable; but our simulated system would only be as unpredictable as a human normally is -- it would generate plausible behaviour, in the same way that we always generate plausible behaviour despite the presence of low-level noise.)

Of course you reply that this is only a digital *simulation*. But this is enough to allow, in principle, computational success in a Turing Test. As follows. The system that we simulate consists of a closed room, with a human sitting in front of a communication terminal. The only inputs to this system are occasional discrete signals, which correspond to certain words appearing on the terminal. (These, in case you haven't worked it out, correspond to the questions that the computational Turing Testee might have to answer.) We can simulate the appearance of these words on the terminal screen, the entry of information into the human's brain, the brain processes, and the consequent motor movement on the part of the human -- all these are simply physical processes obeying the laws of physics. In particular, we can simulate the typing actions of the human, which lead to certain discrete signals being output from the room, corresponding to the *answers* given by the human.

Now, getting a computer to pass the Turing Test is easy. In our simulation, we take the *real* questions which are being asked by a questioner, and convert these to the form of the simulated input signals into the simulated room. We allow our simulated room to proceed, until it eventually comes up with simulated output signals. We then deconvert these (still digitally) into the words that we want the computer to present as its answer. Thus, the computer can produce behaviour equally as good as that of a human on the Test (for it is a human that is simulated).

This may seem complex, but take a moment to look. It's actually very simple. The main point is that to pass the (Linguistic) Turing Test, our sensory functions are inessential (though you claim otherwise). The only input required is a little visual input, and this can be simulated if we simulate an entire system. You occasionally claim that functions internal to the brain are inherently sensory, and maybe analog, but they are physical systems and thus simulable. (Unless you want to claim that they are analog to an indefinitely deep level; a claim which has no supporting evidence, and is incidentally made irrelevant by the noise provided by quantun mechanics.)

Note that I make no claims as to the metaphysical significance of this system. Certainly I believe that this, as a simulation of a human, will have all the mentality of a human, but this claim is not relevant here. I make claims only about the system's behaviour, about its *function*. It clearly *can* pass the Turing Test, computation or no computation. Therefore I hold that your argument that a digital computation could never, even in principle, pass the Turing Test is simply false. You seem to frequently throw this argument in as a back-up claim, without evidence. Admittedly it is not central to your arguments, but it certainly helps them. I would ask you, in virtue of this demonstration to cease making such claims.

- [SH: Please follow this carefully; it will prevent further - misunderstandings: First, I have never ARGUED that a pure symbol - cruncher could not pass the TT in principle, i.e., that this is - logically impossible: In principle, since there's nothing going on in - the TT except symbols in and symbols out, it seems logically possible - that there should be nothing going on in between the I/O except symbol - crunching either. This is explicitly stated in "Minds, Machines and - Searle." But what I added there was that I did not believe this could - be accomplished in practise, that I know it HASN'T been accomplished in - practise, and that, because of the Symbol Grounding Problem, - successfully passing the TT would probably have to draw on the - functional capabilties needed to pass the TTT, and hence that - whatever could pass the TT would also already have to be able to pass - the TTT -- and THAT, as I have shown, CANNOT be all symbolic, because - of the sensory projection at the very least (transducer function), - and probably a lot more like it too. - - Now, I have no doubt that any process can be simulated symbolically (I - accept Church's thesis). So I agree that a furnace, an airplane and - a robot capable of passing the TTT can each be symbolically simulated. - What I deny is that the symbolic simulation alone can heat, fly or - pass the TTT -- for roughly the same reasons: each draws on - essentially nonsymbolic functions (heating, flying, transduction) - that can be simulated by but not implemented as a pure symbol - cruncher. Now here's what your proposal amounts to: - - A symbolic simulation of a robot that (if implemented as a robot) - could pass the TTT cannot itself pass the TTT (for the reasons - already noted). Can it, however, pass the TT? Is the missing - wherewithal for passing the TT by symbol crunching alone adequately - supplied by a symbolic simulation of a robot which, implemented, would - have had the REAL wherewithal to pass the TTT? - - My answer is that I don't know, just as I said I didn't know whether a - symbol cruncher that was designed ONLY to pass the TT could in fact - pass the TT (but I doubt it). In both cases, symbols alone would be - enough in principle, because there’s no nonsymbolic function directly - involved in the TT task (symbols in, symbols out). But what remains - moot is whether or not success on the symbols-only task would have to - draw ESSENTIALLY on nonsymbolic functions (such as heating, flying or - transduction), or could do just as well with simulations of them. - - Who knows? The only intuitions I have are the same as yours: Maybe - to talk coherently about cats and mats I have to have (or better, I - have to BE) analog sensory projections of them, and feature detectors - for picking out the analog invariants. Now obviously these can all be - digitized and symbolized. For successful passage of the TT, how - important is it that these BE in analog sensory form, rather than - symbols? - - But note that the issue is NOT just discreteness vs continuity. - Consider, for example, a "mental" furnace: How important is heat - transduction (or whatever the proper word is, "convection," - perhaps), as opposed to merely a long description of the properties - of heat transduction, for being able to TALK coherently about - "furnaces" and "heat" as we do? Without any homuncularity - whatsoever, it could be the case that, just as a picture is worth not - just a thousand words, but an infinity of them, a sensory analog may - be worth more than an infinity of symbols. For BEING the grounded, - feature-detecting sensory surface (and analog transformations of it), - rather than just being ABOUT such a surface, may make it possible to - anticipate an infinity of contigencies that a finite description - could not anticipate, and making coherent discourse may depend on - being able to appreciate those contingencies. - - By the way, to capture all the causal capacity of a furnace or - airplane in differential equations it seems to me you have to encode - not only the furnace or airplane, but all the rest of the universe in - all of its potential interactions with the model as well. A tall - order for symbols, but something the analog can do in short order by - just BEING what it is, rather than being ABOUT it. SH]

(4) Almost everything has been said that has needed to be said. I rather like your idea of a "Total Turing Test", as a tougher version of the Turing Test. I don’t think that it leads to a deep metaphysical distinction, though. In fact, a variation of the above argument can easily be used to show that a digital computer could, in principle, pass the Total Turing Test, as long as it was equipped with appropriate sensory transducers for vision, hearing, motor movements, and so on. Such transducers would be needed *only* for input/output functions; there is no validity to your claim that such mechanisms are needed "all the way in" inside the brain. Insofar as these mechanisms are not needed for input/output, they are physical systems which can be simulated by an internal processor. Thus, your reply to Searle becomes simply a variation on the Robot Reply: equip the system with appropriate I/O functions and it might pass. The above argument shows that a computer plus I/O could pass a Total Turing Test. Given your professed view that a TTT-passing entity would possess true mentality, this starts to look rather good for computers.

- [SH: I agree that everything’s been said, including why I don’t think - the symbol-cruncher plus peripherals hookup (as in the "robot reply") - could pass the TTT, why the TTT is not just "tougher," but may draw - directly on the nonsymbolic functions that the TT needs to draw on - too, and why the mind may be no more a symbol-cruncher with I/O than - an airplane or furnace is, but rather, like a plane or furnace, - nonsymbolic "all the way in." "Some functions should not mean, but - be." -- Stevan Harnad]

From: Richard Yee Subject: Further clarification of our positions?

From: Richard Yee (yee@cs.umass.edu) To: Stevan Harnad (harnad@clarity.princetion.edu) Cc: Thomas H. Hildebrandt (delta@csl.ncsu.edu)

Stevan,

This is a delayed response to your comments on my earlier posting on Connectionists. Feel free to forward portions of this with your comments for general discussion, but I would ask that the portions be left largely intact; I felt that some of my positions did not come through clearly in your reply last time.

I would like to clarify some points concerning the Chinese Room discussion. In what follows, I describe the situation as I see it. There are 8 points followed by a discussion. I was curious as to where you would agree and disagree.

Best regards, Richard

----

(1) The Chinese-Room question is whether the input symbols, Chinese characters, are (or could be) *understood* according to the scenario given by Searle. In (4) below, I describe this question in more detail and refer to it as "Q1".

(2) The Church-Turing Thesis (computation = algorithm = Turing machine) is not in question. It is assumed to be true.

(3) Ultimately, the question raised by the Chinese Room argument is whether "understanding" is a computation. By the Church-Turing thesis, understanding is a computation IF AND ONLY IF it is achievable by some finitely describable algorithm which is equivalent to saying that it can be implemented as a Turing machine (TM) (e.g., see [1]).

To use your flying analogy, this is the question of whether a TM can really fly with respect to "understanding" as opposed to merely simulating flight. If understanding is a computation, then TM’s *have the right stuff*.

(4) Searle’s definition of "the understanding of a symbol" describes a valid and interesting phenomenon, and it raises interesting issues for cognition and computation.

He states that he does not mean to probe deep philosophical notions of understanding. The understanding of a symbol that he is after, simply means recognizing not only the symbol’s *form* but also its *content*. For example, if X is a typical adult’s knowledge of what a hamburger is, and Sn(X) is system number n’s symbolic representation corresponding to this knowledge, then "understanding symbol Sn(X)" means recognizing that Sn(X) refers to X---in this case, a hamburger. That is:

X = A typical adult’s knowledge of what a hamburger is Sn(X) = System #n’s symbolic representation referring to X

Therefore: Form (Sn(X)) = Sn(X) Content (Sn(X)) = X

For example, we might have: S1(X) = HAMBURGER S1 is English (written words). S2(X) = "HAN-BAO-BAO" S2 is Chinese (written characters). S3(X) = (SANDWICH BREAD: BUN, (OPTION SESAME-SEEDS) FILLER: BEEF-PATTY EXTRAS: PICKLES, MUSTARD, KETCHUP, (IF CHEESE THEN (REPLACE "HAM-" "CHEESE-"))) S4(X) = G0037 S5(X) = a picture of a hamburger etc., etc.

We ask whether the following is possible inside the Chinese Room according to the scenario proposed by Searle:

(Q1) Understand (S2(X)) = Content (S2(X)) = X

(5) The Turing Test (TT) alone (observation of I/O behavior only) is not capable of supporting the conclusion that Chinese is being understood inside the room because this is a question about the process by which the observed behavior arises. The process that is being used to determine outputs is hidden inside the room; hence, an external observer can NEVER be justified in concluding that Q1 is DEFINITELY occurring in the room, regardless of the I/O behavior.

(6) The person in the room is aware of the form of the input symbols, but never knows of their content. Given S2(X), the person's behavior may be affected by his knowledge of the form: S2(X), but he is never conscious of the fact that the content of S2(X) is X. Thus, his behavior can never be DIRECTLY affected by X. This leads you and Searle to the following:

Assertion (6):

Premise: The person never figures out for himself that Content (S2(X)) is X, i.e., the person never performs Q1 directly for himself.

Conclusion: Q1, the understanding of the *content* (meaning) of the Chinese characters, definitely DOES NOT occur inside the room.

The reason for reaching this conclusion is that the person's awareness is the ONLY potential locus of understanding in the system.

(7) Presuming, for sake of further discussion, that the understanding of Chinese input characters were occurring inside the room, an additional question that you have raised is whether this understanding is being *experienced* by some self-conscious entity other than the person ("Where was this understanding last Tuesday?", etc.). This goes beyond the notion of understanding proposed by Searle: the mapping of formal symbols to their content. It is asking for a *meta-level* understanding; namely, it requires the understanding THAT the content of input symbols is understood:

Understand ("Understand (S2(X)) = X") = Understand ("Chinese symbols are understood")

This knowledge implies a self-conscious understand-ER: "I" Understand that: ("Chinese symbols are understood---BY ME") Despite his qualifications, I suspect that Searle, too, would not allow one to claim that Chinese is *really* being understood without this meta-level of understanding. Nevertheless, this level of understanding not what he requires in his scenario.

(I would argue that there is no reason that such an understanding is not also computable, but this is a separate, though related, question. I do not wish to pursue it further here since I feel that it is a complicating factor.)

(8) You and Searle appear to be lead to the following:

A Turing machine (TM) is not capable of treating its input symbols any differently from the way in which the person in the room treats the Chinese characters. Therefore, a TM can never understand its input symbols, and, consequently:

Understanding is not computable.

Discussion:

I do not claim that Chinese is being understood in the room. As pointed out in (5), I do not believe that I/O behavior provides any basis for such a claim. I agree with you and Searle that, as in (2) above, understanding requires that the input symbols be recognized for their content (intentionality, grounding) rather than being treated only with regard to their form. Also, I agree with the premise of Assertion (6): (we can assume that) the person never understands the meaning of the Chinese characters; he never recognizes their content. The main point on which we differ seems to be the conclusion of (6): the content of the Chinese characters is DEFINITELY NOT being recognized. We disagree on this, I believe, because the theory of computation shows that the person's awareness is not the only place where we must look for a potential understanding of the input symbols.

If we are to conclude that the understanding of the input symbols is not being computed in the room, we must consider BOTH of the computations that are occurring. As pointed out in my earlier posting on Connectionists, the person is acting as a Universal Turing machine (UTM). He produces one computation which is the process of applying the rules (memorized or not) to the input symbols. The person's computation, rule application, is the computation that everyone talks about, and this is the computation which I agree does not entail the understanding of the Chinese symbols. In other words, the symbols' forms are NOT mapped to their contents (meanings) in this first computation.

The SECOND COMPUTATION is not as obvious as the first, but its existence is not a matter of conjecture or philosophical argumentation. It is a mathematical fact, no more or less valid than, say, the Pythagorean Theorem. It is simply the case that a UTM's computation produces the computation of another, different, TM. The person's computation produces a second computation that acts on the inputs. We are not told by Searle exactly what the second computation specified by the rules is, but it is this computation that is completely responsible for the input-output behavior of the room. It cannot occur if the person does not carry out his computation correctly, but it is not the same thing as his own computational process.

Thus, we cannot conclude that the answer to our fundamental question of whether understanding is a computation is "no". If it were the case that understanding were computatable, then it could be described by some algorithm. This algorithm could be given to the person in the form of rules which, when carried out, produces the understanding of the input symbols (real understanding, not "simulated"). The input symbols could be mapped to their content (within the second computation) via the actions performed by the person. The person's awareness need only involve rule application (the first computation). (To wax metaphysical, one might say that the physical system of which the person is a part operates as a brain producing a computation that, like a mind, processes the Chinese inputs. We cannot say whether this "mind" understands the Chinese characters or not because we don't have enough information as to how it is working: e.g., how it was formed, etc. It might, however, truly understand the Chinese even though the "brain system" clearly does not.) Since this scenario is completely consistent with the one described by Searle, we cannot conlude from his argument that no algorithm exists for mapping the content of the Chinese symbols to their meanings. There is no reason why a TM could not map the input symbols that it receives into their content and process those inputs symbols accordingly.

>>> Conclusion:

In sum, I agree that treating input symbols as formal objects only, may be able to simulate understanding from the point of view of an external observer, but that this surely *does not* mean that the system understands the input symbols. The problem we have is that, there are TWO Turing machines computing away inside the Chinese room, and Searle has only succeeded in showing that ONE of them is treating the input symbols as formal objects only. However, there is, absolutely no reason that a TM MUST treat its inputs as solely formal symbols. It is possible, I maintain, for a TM to interpret its input symbols with respect to their content, and in the Chinese Room this MIGHT be happening within the computation of the second TM that is processing the inputs...

The Turing machine described by the person's rules may REALLY be able to fly (understand) when its computation is set in motion by the person's actions.

References [1] Lewis, Harry R. and Christos H. Papadimitriou. Elements of the Theory of Computation. Prentice-Hall, Englewood Cliffs, NJ. 1981. (in particular, pp 222-224, pp 258-262).

ON WHAT DOES AND DOES NOT FOLLOW FROM CHURCH'S THESIS

To: Richard Yee From: Stevan Harnad

Dear Richard,

Thank you for your follow-up. I am not circulating it to my Symbol Grounding list because I think it merely retreads ground we have already covered before, and at too great a length. If you wish to circulate it yourself, you may, if you wish, append this message as my reply. This was your summary:

> I agree that treating input symbols as formal objects only may
> be able to simulate understanding from the point of view of an external
> observer, but that this surely *does not* mean that the system
> understands the input symbols. The problem we have is that there are
> TWO Turing machines computing away inside the Chinese room, and Searle
> has only succeeded in showing that ONE of them is treating the input
> symbols as formal objects only. However, there is, absolutely no reason
> that a TM MUST treat its inputs as solely formal symbols. It is
> possible, I maintain, for a TM to interpret its input symbols with

> respect to their content, and in the Chinese Room this MIGHT be
> happening within the computation of the second TM that is processing
> the inputs...
>
> The Turing machine described by the person’s rules may REALLY be able
> to fly (understand) when its computation is set in motion by the person’s
> actions.

You say there is not one but two Turing Machines in the Chinese Room. So be it. There’s still only symbol manipulation going on in there, and Searle (having memorized all the symbols and manipulation rules) is the only one doing it, and hence the only possible subject of the understanding; and he does not understand. That’s all there is to it. I’m certainly not prepared to believe in extra, phantom subjects of understanding under such conditions -- and certainly not because of an abstract construal of Church’s Thesis.

I can’t follow your point [not excerpted here] about understanding vs. knowing you understand. There’s no need to complicate things: Searle is not understanding Chinese in the Chinese room, and he knows he’s not understanding it, and that’s all there is to it. Again, there’s nothing about the circumstances of memorizing a lot of rules for manipulating meaningless symbols on the basis of their shapes that should make me believe that Searle now does understands Chinese, but without realizing it. I don’t believe you can understand a language without realizing it, i.e., while thinking instead that all you’re doing is manipulating meaningless symbols. I think that under those conditions all you are doing is manipulating meaningless symbols.

I also don’t understand your point about "real flying" by a symbol cruncher. Are you saying a computer CAN really fly? Or just that a plane is "equivalent" to (i.e., symbolically simulable by) a computer? The latter is irrelevant: There’s no flying going on in a pure symbol cruncher, just symbol crunching. Ditto for thinking, on all the arguments I’ve heard till now.

I also don’t understand your point about how to get from content to form via "interpretation." Whose interpretation? Certainly not Searle’s. I see nothing but symbols and symbol manipulation going on in the Chinese Room. You’re saying [in a passage I have not excerpted]: Let the meaning of symbol "S" be "X." But "X" looks like just another symbol to me. Where do you break out of the circle of symbols? Apparently you don’t; you simply succumb to the hall of mirrors created by projecting your own interpretations onto those symbols.

Another thing I don’t understand is what you mean by the "actions" of the person. Do you mean the symbol manipulations themselves? Or do you mean actions on objects in the world (which would be another ballgame: the TTT instead of the TT, and no longer just symbol crunching)? I see nothing in the "action" of manipulating meaningless symbols that comes alive as understanding. Do you? If so, you ought to be able to give me the flavor of it -- not just more abstractions about the consequences of Church’s thesis. If Church’s thesis does not entail that a pure symbol cruncher, though computationally equivalent to an airplane that flies, can itself fly, then it does not entail that a pure symbol cruncher, though computationally equivalent to a device (e.g., a brain) that thinks, can itself think.

Please, if you reply again, try to be brief, and stay down to earth, as I do. Don’t claim the existence of extra entities because of some way an abstract equivalence relation can be interpreted when there are no other entities in sight. And don’t claim something can really fly when that’s not what you really mean. And above all, try not to lose sight of the difference between a meaningless symbol token such as "X" and the meaning or content it can be systematically interpreted as having. Until you give down-to-earth evidence to the contrary, that content cannot just consist of still more symbols, or the "action" of manipulating them.

Stevan Harnad

COG SCI VS. SCI FI

Michael G Dyer wrote:

> [You] seem incapable of seeing how Searle's brain could be used to
> implement a distinct virtual system (even though that's done all the
> time on the simpler von Neumann architectures).
>
> Here's a test that CAN ACTUALLY BE CARRIED OUT, and that might shed
> some REAL light on the issue of qualia. This researcher has turned
> people's backs into a primitive kind of "retina" i.e. a bunch of pins
> one wears on one's back push in to make some kind of "shadow image"
> from info coming into a camera... [it] gives blind people some idea of
> what untouchable things... are like... here's a testable question:
> after wearing the thing for days, will one "forget" that it's back pins
> and start "seeing" directly? or will one keep thinking that one's not
> really "seeing", but "just" interpreting back pins?

The test is completely irrelevant to any of the issues under discussion. Predictions about what sorts of sensory transformations can be made by an already grounded system have nothing whatsoever to do with what an ungrounded symbol cruncher can or can't do. And I would suggest that you resist projecting such fantasies and analogies onto your von Neumann architecture -- virtual or otherwise -- because it just puts you in a hall of mirrors in which you can no longer tell that the only source of the projection is YOU.

> You seem hung up on transducers. First, a computer is a transducer
> because it is physical and operates by moving actual physical things
> around.

You have apparently not understood the transducer argument. It is not the transduction that is involved in implementing a symbol cruncher that is at issue but the transduction in the sensory systems of a robot capable of passing the TTT. The argument was that whereas Searle CAN perform all the functions of a symbol cruncher (by "being" its transducers/effectors for manipulating symbol-tokens) while demonstrably not experiencing the mental state of understanding, he cannot perform all the functions of an implemented robot without experiencing the mental state of, say, seeing; for his only two available options are either (1) to use only the output from the robot's sensory input transducers (in which case he is not performing all of the system's functions and therefore rightly vulnerable to the "systems reply") or (2) to BE its transducers, in which case he would indeed be seeing. Either way, the Chinese Room Argument fails for a grounded robot and the TTT whereas it succeeds for an ungrounded symbol cruncher and the TT.

> Consider phantom pain in severed limbs...

Irrelevant to an ungrounded system, same reasoning as above.

> we NEVER experience the sensory world, we just experience the signals
> from our own transducers. But even THAT is not true, we actually just
> experience the structure of our own brains! (I.e. each part of our
> brain just gets signals from the transducers and other parts of the
> brain). We talk AS IF we are experiencing the world because it's much
> easier to talk that way.

Look, all of this elementary philosophising about what we're really experiencing is fine -- some philosophers say it's objects out there, others say its things going on in our heads, etc. -- but irrelevant. No matter what your fix on the mind/body problem happens to be, on what's been said and shown so far, there's no justification for projecting ANY experience of ANYTHING (of objects, inputs, transducers, internal structure, or what have you) onto a symbol cruncher.

> The only reason we feel that the world makes sense is because there is
> COHERENCY in what our different sensors tell us as we move about in the
> world that we hypothesize exists. Those whose senses aren't in decent
> enough correspondence with the hypothesized "real world out there"
> (i.e. don't see there's a car coming) don't survive to pass on their
> 'defective' neural circuitry.
>
> The AI working hypothesis is that the specific transducer does not
> matter, only the PATTERNS it produces and the CAUSAL COHERENCY between
> the transducer input patterns, their interpretations in the brain, and
> the motions of the organism in the environment (whether consistent
> enough or not wrt survival).

There seems to be a little bit of conflation here between coherence theory and correspondence theory, which are usually regarded as alternatives, plus an added dash of causality, but no matter. It has always been the dream of top-down symbolic functionalism that the real cognitive work will be done by the symbol cruncher, and then it's just a matter of hooking it up to the world, through trivial transducers/effectors, "in the right way" ("correspondingly" and/or "coherently," and in either instance "causally").

Suffice it to say that my grounding arguments call this dream into question. If these arguments are sound, the problem of hooking up to the world "in the right way" will prove to be where most of the substantive problems of cognitive modeling -- i.e., generating TTT performance capacity -- really reside, with much of the solution consisting of nonsymbolic function, and the residual symbol crunching turning out to be what is comparatively trivial. And as I've said over and over, the resulting grounded robot may well turn out to be a dedicated, hybrid symbolic/nonsymbolic system in which there is NO isolable symbol crunching going on at all -- and certainly none that could be described as simply a symbol cruncher hooked up to the world "in the right way" through its transducers/effectors.

If all this can be anticipated by an ingenious theorist in advance, then of course it can all be simulated symbolically -- the robot, its innards, its input and its output. That simulation will have all your coherent "patterns," but it won't think, any more than a simulated airplane can fly or a simulated furnace can heat. All three, respectively, will simply be INTERPRETABLE as thinking, flying, and heating. To really think, fly and heat, they will have to be grounded in their REAL I/O by being implemented as the respective nonsymbolic systems that are here merely being simulated by the symbol cruncher, namely, a TTT-passing robot, a plane and a furnace. And, as I said, in none of the three cases will this grounding amount to simply "hooking up" the symbol cruncher to real I/O "in the right way."

(Try to keep flying and heating in mind as examples before you come back with your hall of mirrors again: Merely being interpretable-as flying or heating is not the same as actually flying or heating; the same is true of thinking. Any "interpretation" has to be made intrinsic to the system itself by being grounded in its performance in the real world; otherwise it's just a reflection of what we ourselves project onto it, an illusory mirror-image.)

> The Searle working hypothesis seems to be that the material of the
> transducer does fundamentally matter. All experience in computer
> science supports the AI working hypothesis (e.g. you can get similar
> behavior out of machines made of completely different transducers).

Incorrect. It's not the material that matters to Searle, as he's said over and over; it's the causal powers. He is not a sceptic about synthetic brains, for example, as long as they have the causal powers of real brains. Pure symbol crunchers do not. And their inner lights don't magically go on if you hook 'em up to transducers either.

> changing the transducer [by] placing a prism in front of the eyes...
> after a while, the subject again sees the world as right-side up...
> you replace my hand with an artificial membrane I should not notice the
> difference, as long as the same patterns are sent to my brain as the
> original hand sent. The same argument can be made for replacing
> selected neurons in the brain, ... until the entire brain is replaced
> by artificial neurons.

Fine, but you're back in your hall of mirrors again: Replacing something like a transducer with something that has the same causal powers is not the same thing as simulating it on a symbol cruncher and interpreting it as if it had the same causal powers! What if most of what you're replacing as you go through this brain prosthesis exercise turns out to be components with nonsymbolic causal powers like transduction? Their causal equivalents will still have to be able to do real transduction. Recall the analogy with airplanes and furnaces: Replace their parts with functional equivalents and you still don't end up with just a symbol cruncher hooked up to I/O transducers. Or if you do, precious little of the work is being done by the symbol cruncher...

> Suppose there were creatures (called "Photonians") made of reflected
> laser light... "photocircuitry" on the same order of complexity (in
> terms of computations) as our neural circuitry... "Photo-Harnad"...
> "Photo-Searle"... arguing that, although humans act like they have
> intentionality, they really don't, because their brains are not made of
> photocircuitry and light acts in fundamentally different ways than

> human neurochemistry... save the Earth...

So many things are conflated in this sci-fi fantasy that it's almost impossible to sort them out. First, can we by now set aside once and for all the red herring about scepticism concerning physical materials, as long as they have the requisite causal powers, please? Now, it's not clear to me how you imagine that creatures made only of reflected laser light could pass the TTT (e.g., how do they manage to see and manipulate objects as we do?), but if they have the requisite causal powers to do so, I have no reason for scepticism.

Attributing magical powers to materials in a sci-fi fantasy, however, is not a very rigorous form of argument. Suppose rocks had the power to think, suppose gas had the power to feel, and suppose the Atlantic Ocean had the power to pass the TTT. The Chinese Room Argument certainly couldn't demonstrate otherwise. But the fact that nothing is immune from a fanciful sci-fi interpretation does not make the fiction into reality. And so it is with the overinterpretation of the powers of symbol crunching (be it ever so "computationally complex") -- which you like to rescue from the Chinese Room Argument by projecting a multiple-personality fantasy onto Searle. In the symbolic hall of mirrors, anything is possible...

Stevan Harnad

THE LOGIC OF THE TRANSDUCER COUNTERARGUMENT AND THE TTT

Mike Dyer writes:

> Let me see if I can state YOUR argument in my own words. Here's
> my attempt: >
> 1. In the chinese box version (of the TT), Searle does EVERY aspect
> of the chinese box and does NOT experience that he is understanding
> Chinese, therefore there is no "intentionality".

Correct. By the way, I don't use the word "intentionality" because I don't know what it means. If I'm speaking of understanding, say, Chinese, I just say, "understanding Chinese," and that's what neither Searle nor the symbol cruncher is doing under these conditions. The less mystery, the better, as far as I'm concerned. (Minor point; probably nothing to worry about.)

> 2. In the TTT, Searle must now control a robot that walks and sees,
> etc. To do so, Searle now only has two choices:

No, controlling a robot isn't his mission. It's performing all the functions that are responsible for the robot's performance capacity, with the hope of showing that something (mental) will not be true of him that we were supposing to be true of the robot when it's doing exactly the same thing. In the case of the standard, nonrobotic TT, that (mental) thing was understanding Chinese, as opposed to merely acting in a way that is interpreted by us as understanding Chinese. In my own variant (the transducer counterargument for the TTT), that (mental) thing was seeing, as opposed to just acting in a way that is interpreted by us as seeing. (Another minor point, but one can't be too careful; any loophole leaves room for misunderstanding...)

> (a) he can do, say, matrix manipulations (specified by the
> chinese books) on the vectors being given him (from the photo-
> sensors of the robot) and in this case Searle will NOT feel that he is
> seeing; however Searle is NOT doing ALL aspects of the task,
> since he is NOT simulating the photosensors themselves.

Here is a more substantive point: You already seem to be heading for the reply that transduction is trivial and it's just a matter of hooking up transducers to a symbol cruncher. The photoreceptors basically just generate a huge bit-map, which is then vectorized and ready for the real work, in the form of symbol-crunching. Note that, although I agree that this is a logical possibility in principle (as is the possibility of passing the TT till doomsday by symbol crunching alone), I don't believe it happens that way in practice, either in the brain, or in any system with full TTT-power (just as I believe you'd need TTT-power to pass the TT). The logical point (against Searle) about the immunity of transducer function, however, is independent of this belief of mine and its supporting arguments.

But if you ARE preparing to give me the usual "transduction-is-trivial, just hook the symbol-cruncher to peripherals and it'll still be doing most of the work" argument, bear in mind the possibility (highly probable, I think) that a LOT of the functions required to pass the TTT may be more like transduction (and heating, and flying), i.e., essentially nonsymbolic, rather than symbol crunching. And that very little of what's going on in a grounded, TTT-capable system may end up being symbol manipulation. Grounding is not likely to be a matter of plugging your symbol cruncher into a video camera and suddenly the (mental) lights go on...

But so far, apart from the assumption about bit-maps and vectors, you've still got my logic right.

> or
>
> (b) Searle can use his owns eyes to see with and then take that
> information and do what calculations the books say to
> do, but in that case you claim that Searle WOULD have the
> subjective experience of seeing (moving) etc.

Right. (And of course, I don't think the rest will just be "calculations the books say to do...", but a lot of nonsymbolic and hybrid functions.)

> So, either Searle fails to do ALL aspects of the system (in which case
> it's ok that he doesn't subjectively experience seeing/moving etc,
> since he's failed to do ALL aspects) or he DOES experience seeing
> (since he's using his own eyes etc.)

Correct: Unless he does EVERYTHING, he's open open to the "Systems reply."

> The morals to be drawn from the above are:
>
> (1) that the TTT is MORE than the TT, and

In the sense that it not only calls for more kinds of performance capacity, but it necessarily draws on more kinds of underlying function, e.g., transduction.

> (2) without the photoreceptors (or some other transducer -- from
> actual energy in reality to "informational energy" that computers chinese books> can manipulate) the input vectors being
> manipulated lack a semantics -- since to have a semantics they must
> be grounded in the sensory world (i.e. come about via the
> transducers, i.e. the part Searle can't simulate just by making
> calculations specified in a book).

Well, now it's beginning to sound hirsute again. I don't know what "informational energy" is (and I have a feeling you're headed towards "patterns" that are basically just computational -- and if so, I'm not with you), but WHATEVER it is about transduction that might be special, Searle can't do it, and therefore his Chinese Room Argument cannot be used against the claim that the TTT robot sees in the way it can be used against the claim that the TT system understands.

The LINK between the two mental properties -- seeing and understanding -- according to my grounding theory, happens to be the hypothesis that to understand what a symbol means you have to be able to pick out, with your senses, the object it stands for (in the case of elementary sensory categories, and that the meanings of the rest of the higher-level symbols standing for more abstract objects must be grounded in these primitive sensory meanings).

Now that's just my own grounding hypothesis. The logic of the transducer counterargument is merely that the Searle strategy, though it works for the TT and "understanding" does not work for the TTT and "seeing." My theory tries to pinpoint why, and what the connection between the two might be, but it is important to note that the logic is just that it works for this and not for that.

And of course I don't think successful grounding will just amount to photoreceptors plus books. That may work for for the trivial toy performance we've generated so far, but not for TTT-scale performance capacity.

> (3) so robots that lack sensory transducers, although they act as if
> they have intentionality, really don't, and >
> (4) robots that pass the TTT have intentionality. I have no idea why you put it in this awkward way, so I'm worried about where you're headed: The conclusion as I would have stated is:

(3) So pure symbol crunchers, even TT-scale ones, unlike TTT-scale robots, can't understand anything, as shown by the Chinese Room Argument, even though they can be interpreted as understanding. A TTT-scale robot would be immune to a homologue of Searle's Argument that was directed at seeing rather than understanding.

(4) There is no more (or less) reason for doubting that a TTT-scale robot would have a mind (i.e., would see and understand) than there is for doubting that any other person has a mind (the other-minds problem).

(Except, perhaps, for the "TTTT" -- which, to play it even safer than we do with one another, calls for the full "performance capacity" of not only our bodies, but our brains, and of all the parts of our brains; I happen to think that only those brain powers that are needed to pass the TTT are relevant to having a mind, and hence that most of the hard work will already have been done when we can get ANY system to successfully pass the TTT indistinguishably from the rest of us; the only thing left to do after that would be the fine-tuning, according to me. Searle may disagree, however; and either of us could be right. So nothing essential should ride on this, one way or the other.)

> BEFORE I respond to the above argument, I want to make sure that I
> have YOUR argument RIGHT. so please correct any errors in the
> above (especially the conclusions part)
>
> -- Michael
>

Apart from these caveats, you've got it right.

Stevan

TRAPPED IN THE HERMENEUTIC CIRCLE: Part I [SH.]

From: Dr Michael G Dyer

- [Annotated by me. SH.]

The Turing Test

First, with respect to the chinese box (TT -- standard Turing Test) I will assume that Searle is the one carrying out the computations (logico-symbolic, and/or numerical calculations) specified by the book(s) in the chinese box. To make things more black/white, I want the behavior specified by the chinese books to create a simulation of Minsky's personality, beliefs, arguments etc. So when we pass slips of paper to the chinese box, the conversation we have (in chinese) is with a "system" that can fool us into thinking we're talking with Minsky. (I'm doing this so we can keep the Minsky system straight from the Searle who is helping the Minsky system to come about.)

- [SH: Minsky in CHINESE, I presume you mean, which may make for some - anomalies; so maybe Deng Xiao Ping would be a better alter ego -- but - never mind, let's go on...]

1. Problem with introspection in TT

I have a problem with accepting Searle's introspective experience as being relevant with respect to whether Minksy's intelligence resides (as a system) within the chinese box. Here's my position:

If Searle carries out the chinese-box instructions in his head and does NOT experience being Minsky, then Searle is just a subsystem of the larger system and the fact -- that Searle himself is intelligent enough for us to have a conversation with him -- does NOT eliminate the existence of Minsky, with whom we are also able to have conversations (in chinese).

- [SH: Mike, we're just back to your multiple personality story again. - Forget the "subsystem"/"larger-system" talk for a moment and focus - on what is actually going on: The only one there is Searle, who has - memorized a bunch of rules for manipulating squiggles and squoggles - he doesn't understand, and you're telling me that this has induced - another mind in his head, one that he is not aware of, but that is - itself aware, and understanding the symbols. (It IS aware, isn't it? - I mean, you ARE supposing a SUBJECT for the understanding that is - allegedly going on in Searle's head without his knowledge, right? - Because if you're not -- if this is not the kind of understanding - that has a conscious subject -- then there's no point continuing the - discussion, because that's the only kind of understanding I know or - believe in: literal understanding. Unconscious, subjectless - "understanding" sounds to me like an empty figure of speech.) - - Well, as I said before, I simply do not believe that memorizing a - bunch of meaningless symbol-manipulation rules and then performing - them is the kind of thing that can induce multiple personality. I - think you're projecting a mentalistic fantasy onto Searle -- possibly - because you've gotten into the habit of projecting the same - mentalistic fantasy onto your computer for too long. SH.]

If Searle carries out the chinese-box instructions SO AUTOMATICALLY that he is just in some sense "being Minsky", then we have a paradox of consciousness, since the experience of"being Minsky" is in conflict with the experience of "being Searle" (i.e. having the state of consciousness of being Searle, with his thoughts, beliefs, past memories, reactions to points made in discussions, etc.). So, either Searle is NOT doing the chinese-box instructions THAT automatically, or, to be THAT automatic, I will simply claim that Searle's introspective experience will be to BE Minsky and thus, Searle, while performing these instructions, will have no sense of being Searle.

- [SH: The most logical way to resolve a "paradox" of your own making - (otherwise known as a reductio ad absurdum) is to let go of the - hypothesis that generated the absurdity. That hypothesis was: - "Thinking is symbol crunching." - - Speed and automaticity have nothing to do with it. There's no reason - WHATSOEVER to believe that if, Searle executed the memorized - instructions more quickly or automatically, then anything as radical - as a new mind would result. All real evidence suggests instead that - as a task becomes automatic, it becomes more mindless, not that it - gives rise to another mind... (Consult the Schneider & Shiffrin - automaticity literature). Doesn't it bother you that to support my - position I just have to resort to ordinary data and common sense, - whereas to support yours you must must have recourse to increasingly - far-fetched fantasies? - - I don't think the standard Churchland reply (that common sense has - erred before, in the history of science) is at all relevant here, by - the way, partly because we're only talking about thought experiments - and Sci Fi, and partly because that counterargument has the base-rate - and the conditional probabilities all wrong: It's like saying that - the (alleged) fact that Einstein was a slow learner somehow makes my - little Johnny, likewise a slow learner, into a potential Einstein! - SH.]

Notice that I am allowing Searle to do EVERY operation of what is needed to make the box (the system) behave as Minsky (in the TT). Now I know you laugh at the idea of "split personalities", but in computer science, at least, we have great experience with the I/O capabilities of various subsystems being TOTALLY different than the I/O of the"emergent" system.

- [SH: It's not the I/O capabilities that I'm laughing at (they're - being accepted ex hypothesi, remember?): It's the personality (mind) - that you're projecting onto them. That's a hermeneutic habit you'd do - well to try to set aside for the sake of this argument, because it is - in fact ITS validity that's at issue here in the first place, as we - test the hypothesis that "thinking is symbol crunching."]

The reason people tend to (naively) accept Searle's introspective argument that he doesn't understand chinese (and also doesn't experience being Minsky) is that the subjective experience of carrying out instructions can be very different, depending on how automatically those calculations are being carrying out.

- [SH: Are you speaking from personal experience? I mean, I know what - it's like for a conscious skill to become more mindless as it becomes - more automatic, but I've never experienced the eruption of another - mind as a consequence, nor has it been drawn to my attention by - anyone else witnessing my automatized behavior (for according to - your hypothesis, the emergent mind could be one I was not aware of). - Nor have I witnessed anything that looked like this in another - person. Have you had this experience? What's the evidence? Or are you - just referring (yet again) to your OWN hermeneutic experience when - you project understanding onto your computer? If so, we have to - figure out a way to put that pre-emptive conclusion on hold somehow, - so it doesn't settle every possible outcome a priori... - - It also worries me when you talk about what Searle does and does not - understand in the Chinese Room, rather than what he would or would - not understand. A small expository distinction, perhaps, but a - healthy reminder that we are only dealing in hypotheticals here, and - possibly counterfactual ones. SH.]

In the case in which Searle is simulating the Minsky system, people (naively) imagine that Searle is having time to sit there and think thoughts OTHER than what's involved in carrying out the chinese book calculations. In other words, Searle is able to think (as Searle) "gee, doing all these matrix multiplications and logical operations is a bore" WHILE Searle is carrying them out. So even while the RESULT of these operations creates a Minsky system (with which we have conversations), the Searle who is doing it is NOT experiencing BEING Minsky (or understanding Chinese). If this is the case, however, I will simply claim that there are two systems and we can't ask Searle to introspect in order to find out what it's like to be Minsky. We have to ASK THE MINSKY SYSTEM what it's like to be Minsky.

- [SH: Let me see if I have this right: If Searle is kept really busy - following the instructions then what? Then he must have two minds? - Why was that again? And if he's less busy, then what? I'm trying to - pin down the necessary and sufficient conditions for the induction - of multiple personality disorder, according to you. It has something - to do with memorizing meaningless rules, but I can't quite get - whether it's induced by doing them fast or slow, automatically or - reflectively, with boredom or gusto...]

Now let's again imagine that Searle does the calculations SO AUTOMATICALLY that he ceases to be aware of it, and not only does Searle do these calculations, he does ONLY these calculations -- that is, his entire mind is being used in doing these calculations. NOW why should I accept his claim that he (Searle) won't experience being Minsky? There's NOTHING ELSE GOING ON in his head OTHER THAN those calculations involved in creating Minsky. While these calculations are going on, there is now ONLY MINSKY residing in Searle's brain. If we allow Searle to stop doing the computations needed to create Minsky, then now we are talking to Searle and again, he will deny being Minsky (or understanding Chinese).

- [SH: Has this ever happened to you? I mean, have you ever done a - complex, all-consuming task so automatically that you went blank? - And then, when you stopped doing it and came to, someone told you: - "Hey, do you realize you've been babbling on in Chinese about being - this professor from MIT with a very unChinese sounding name..." - - Mike, tell me why I should go along with any of these fantasies: - Do you have evidence or an argument? Or are you just bent on

- saving prior assumptions at any counterfactual cost? SH.]

So I claim there are two possibilities:

1. Searle has time"to be Searle" while carrying out "Minsky- forming" computations, so I can have different conversations with Searle (than I will have with Minsky) -- i.e. two personalities within one system.

2. Searle does automatically and ONLY the Minsky-forming computations. During this time there is only one system operating, and Searle does not exist during this time period, only Minsky. So if the real Searle claims that he would not experience being Minsky, I simply will disagree with him - [SH: Sounds like multiple personality both ways: simultaneous - in one case, sequential in the other -- but (to me) absurd either - way, based on all the available human evidence about multiple - personality disorder, rule-following and automaticity -- not to - mention common sense and everyday experience.]

(footnote: notice that there are actually many possibilities inbetween, depending on how much of the Searle-forming computations are allowed to occur while the Minsky-forming computations are occurring. We will get many strange forms of consciousness -- beyond my ability to imagine!)

- [SH: Your imagination has ALREADY left me far behind...]

So, from my point of view, Searle's argument never dealt any kind of blow to the "systems reply" and Searle's argument still relies on confusing levels of systems and/or subsystems, and the extent to which personality-forming computations are going on at a given time.

- [SH: No blow will ever be dealt to the systems reply (which is just - a restatement of the "thinking is symbol crunching" hypothesis that - Searle was testing in the first place) as long as unconstrained - fantasies can be spun out at will to protect it from any contrary - evidence or reasoning.]

My own brain right now is taken up (100% I think) with Dyer-forming computations. I would hate to have Searle deny me MY experience of being conscious because Searle can imagine doing the computations my brain is doing right now and then claim that HE would still not feel like ME. If I ACT as if I am conscious and can argue it cogently and endlessly, then you should accept me as BEING conscious. The same applies for the "Minsky System".

- [SH: BOY are you in it deep! You mean you think that in simulating - Minsky, Searle would be BEING Minsky? and at Minsky's expense, yet? - What do you suppose Marvin (the one back at MIT while all this is - going on, remember?) would have to say about all these goings - on? "Hey, quit messing with my mind!"? You have given the power of - the imagination a new dimension, in more ways than one... - - But, as I've already said, if the hypothesis that "thinking is symbol - crunching" (henceforth "T = SC") is false, then your brain is not - just doing symbol crunching, and hence it's safely immune to Searle's - thought experiment. Being trapped in a hermeneutic circle of your own - making, you seem unable to understand this simple logical point. SH.]

(I've earlier mentioned the chauvinistic danger of denying intelligence to something even though it behaves as if it is intelligent).

- [SH: There's dangers either way you go (because of the other-minds - problem), but you seem to have gone off the deep end in the other - direction.]

2. Problem with "simulation not= real-thing" argument.

This argument has been stated by Searle when he said that "intentionality is like digestion" and when you (Harnad) stated that a simulation on a computer of an airplane flying is NOT the same as a real airplane flying. Sure, there is a difference, and there are lots of phenomena of this sort. E.g. a simulation of information about the movements and destructive damage of a tornado is not the same as the actual tornado, etc.

Whether one accepts this argument or not, unfortunately, depends on one's preexisting assumptions about the nature of intelligence (intentionality etc.). So we end up in circularities in everyone's arguments (including my own).

- [SH: No, no: YOU end up with circularity. If *I* had been the one - holding the prior hypothesis that thinking is symbol crunching and I - was then confronted with evidence or arguments that it wasn't, I - hope I'd give it up or at least revise the hypothesis -- rather than - resorting to Sci Fi to save it, justifying the Sci Fi with the fact - that, after all, the hypothesis is true! Now THAT's circularity.]

For example, I ALREADY believe that intelligence is manifest in the real world by a real machine that moves real electrons (or photons, or chemicals, etc.) in ways specified by the causal architecture of the real machine. As long as the patterns of energy (i.e. the information) are in the same causal relations to one another, it doesn't matter (for me) whether or not one machine is a universal simulator, simulating, say, a neural network, or if the machine is a hardwired neural network.

- [SH: All you KNOW is that your brain has a mind. You don't know what - it's doing to have a mind (no one does). You have only HYPOTHESIZED - that the only relevant thing it's doing is symbol crunching. On the - face of it, "moving electrons" and "moving symbols" are not - necessarily the same thing). So when you're faced with evidence that - symbol crunching alone is not enough to have a mind, what you OUGHT - to do is revise your hypothesis rather than reiterate what you - already believe. It's the validity of your prior beliefs that's at - issue!]

So again we have an impasse. SINCE I postulate ahead of time that intelligence arises from information processing, any machine capable of processing information that exhibits intelligence is going to be acceptable to me as HAVING intelligence. However, those who postulate that there is MORE than information processing are going to accept the "flying not= simulation of information about flying" argument. While I accept that "flying not= simulations of flying", I am NOT affected by this argument since I believe that intelligence is not anything more than information processing and a computer is NOT simulating information processing, it's DOING IT in the real world (with real energy patterns that embody the information it's manipulating).

- [SH: Rechristen your hypothesis a "postulate," repeat it - enough times, and you're effectively immune to contrary evidence or - argument. I have an even better idea: Why don't you just DEFINE - thinking, understanding, etc. as symbol crunching, and that'll be the - end of it? - - By the way, that flying is not simulated flying is a fact, not an - argument. It's not even an argument (by analogy) about thinking, - particularly. It's Searle's argument that's the argument, and the - flying analogy is just meant to remind you (in case you've forgotten) - that there are more functions under the sun than symbol crunching, - and that thinking, like flying, seems to have turned out to be one of - them. SH.]

So, the "simulation of flying" argument doesn't convince me, since I think that, while flying is more than moving about INFORMATION about flying, I think that intelligence is ONLY moving information around, so a computer is a perfect device to BE intelligent, and not just to "simulate" being intelligent.

- [SH: Mike, I don't know what you mean to mean by "intelligence," - but if it's anything mental, I beg to differ. And this - "information-moving" business is again just a reiteration of - the T = SC hypothesis, gussied up in different terms. SH]

[End of part I (Turing Test). Part 2 (Total Turing Test) follows.]

TRAPPED IN THE HERMENEUTIC CIRCLE: Part II [SH.]

To: Stevan Harnad From: Michael G Dyer

Total Turing Test

Now, with respect to the TTT (Total Turing Test), which includes handling sensory information coming in from the world.

Here, my general position is that grounding symbols is important, but you've elevated the sensory aspect of "symbols" a bit too high. Also, I don't see transducers as being an essential element in intelligence (below I will argue why I feel this way). While I agree that you have found something that Searle cannot simulate (i.e. the photo-receptors or other transducers), I don't think that this is the main way to refute Searle. Why not? because refuting Searle this way leaves one basically agreeing with Searle, i.e. that an (ungrounded) Turing machine can never do more than act AS IF it understands. If one accepts my arguments (above) [Part I], then one can disagree with Searle and still maintain that an (ungrounded) Turing machine, behaving with full human conversational powers, does have "mind".

- [SH: You're quite right that my transducer counterargument does not - save pure symbol crunching. But in Part One you've unfortunately - given no argument WHATSOEVER by way of rejoinder, only fantasies so - counterfactual that they could save anything, from Ptolemy to Special - Creation.]

1. Do "symbols" have a "semantics" or "meaning" if they are not "grounded"?

Now, I want to avoid getting into a lot of problems wrt what these words mean. If we can agree on terminology it will help reduce the amount of (at least gratuitous) disagreement.

There are two directions at which to approach symbols: (1) from abstract, structured, logical relations (normally emphasized in AI) and (2) from sensory experience (normally emphasized in the neurosciences). The issue of the nature of grounding is central here, because that is where we can discuss how these two directions might hook up.

Unfortunately, we lack decent terminology to talk about grounding. For example, X can argue that symbols "lack semantics" if they are not grounded in sensory experience, while at the same time Y can argue that symbols are NOT even symbols unless they capture abstract, structured relationships between patterns that are INVARIANT with respect to sensory experiences.

[SH: That's why I think we should abandon X's and Y's arbitrary stipulations and use the performance criterion that we use every day with one another: The Total Turing Test. According to THAT, sensory grounding will be essential, and primary. The rest is just pre-bout wrangling about what the functional winners in the TTT match are likely to be. I'm betting on a lot of nonsymbolic function.

By the way, systematic semantic interpretability is one of the defining features of a formal symbol system. Formal objects, such as "symbol system" can be defined, and things can be proved about them. But what mind, meaning, thinking and understanding are cannot be settled by definition! They are whatever they are, and it's our business to come up with hypotheses about what functions give rise to them. And then to test them. "Thinking is symbol crunching" (T = SC) was such a hypothesis. It seems to have fallen upon hard times. And on the basis of a THOUGHT-experiment, yet (although its performance score-card is not that impressive either)! So perhaps it was something of a nonstarter to begin with... SH.]

In fact, one reason AI has been successful (to the extent that it has) in modeling aspects of natural language comprehension is that words like "irresponsible" are in some sense easier to handle than simple words, like "move". A word like "irresponsible" has a lot of abstract structure, invariant wrt sensory experiences. It involves a communication, an agreement, the creation of beliefs, a violation (causing a goal failure) etc. Ultimately these must all be grounded in the sensory world, but there are infinitely many ways these groundings could be done and it's the abstract structural relations that make up the essence of "irresponsibility", not any particular sensory experience. On the other hand, a word like "move" is extremely close to the sensory world. All of the inferences involved (causal consequences, enablements, etc) depend intimately on sensory experiences (of what is moving, how, where, when, with what body, etc.).

[SH: I couldn't disagree more. I think concrete vocabulary is primary, and grounded directly in sensorimotor categories, and that the more abstract words are in turn grounded in those concrete ones, and hence in their groundings. The "abstract structure" of the more abstract words ("irresponsibility") consists of their relations to the less abstract words in which they're grounded ("communication," "agreement," "violation," "goal," etc.). It only seems "easier" with the abstractions because you're so far from the ground floor that it looks as if you'll never have to touch bottom -- as if it's all "abstract structure" through and through. That's typically the feeling you get in what I've called the "hermeneutic hall of mirrors," created by projecting your own meanings onto meaningless symbols and their interrelations. What you forget is that, if the symbolic processes are to be mental processes, then their interpretations must be grounded in the symbol system itself, not in you. And a pure symbol system has nothing to "ground" symbols in -- other than more meaningless symbols. (That's the symbolic circle -- what I illustrated in "The Symbol Grounding Problem" with the "Chinese/Chinese Dictionary-Go-Round." The hermeneutic circle in turn arises when you project your own interpretations onto the symbol system and then forget or ignore where they came from in the first place, namely, your own mind. THAT's ungroundedness.)

The natural nonsymbolic place to ground symbols is in the things they stand for -- or, more specifically, in the sensory functions that will pick out the things they stand for. This is what must be done at the level of the lowest level of a grounded system -- the ground floor, so to speak.

All this talk about abstract structure and sensory invariance is fine, but it’s merely a repetition of the defining properties of formal systems. What’s remarkable about certain formal systems is that they are indeed interpretable in ways that fit systematically with other things -- in mathematics, they fit with other formal systems, and perhaps some eternal Platonic verities, and in physics, they fit with the regularities of the natural world. But no one would claim that a symbolic simulation of planetary motion, no matter how well it could be interpreted as fitting real-world regularities, WAS real planetary motion: So somebody seems to be misapplying the lessons of formal modeling here, in claiming that they lead to T = SC inexorably.

At the root of the misunderstanding is the infinity of hermeneutic possibilities afforded by symbols that are interpreted linguistically -- the linguistic Turing Test, which, as I’ve said over and over, is completely equivocal: It’s the equivalent of ASKING a planetary motion simulator to TELL you what it would do in such and such a selected circumstance, rather than setting it to work in the real world on all possible circumstances, as real planetary systems must do. But even more important than that, unlike mathematical functions, which really have to deliver the goods systematically in order to support a coherent interpretation, symbol systems that merely stand for words, or for verbally labeled and expressed "concepts," can give a hermeneutic illusion of coherence and generality by simply being parasitic on our own language and knowledge. As long as you just poke it verbally, your representation of "irresponsibility" looks as if it’s delivering the goods.

Notice that there’s no system that can pass the TT now, merely tiny "toy" fragments that seem to do so only as long as we don’t push them to their limits. But because of the power of natural language, and the compellingness of even a provisional interpretation, our fantasies do the rest for us, and we’re soon hopelessly lost in the hall of mirrors created by our own projections of meaning on these limited and meaningless symbols. Notice that this is NOT like the power of, say, the Peano system to sustain all the (computable) truths of arithmetic: There "Turing Test" the is formal deduction and induction, and the system can pass it all. But with the linguistically interpreted systems we are closer to the position of the ape-language researchers: Once they allowed themselves to "translate" the chimps’ very limited repertoire of signs as English, the sky became the limit.

(Penny Patterson is now circulating fund-raising literature reporting that a captured gorilla, taught sign language, "told" her about how they had killed his mother! All efforts to help and protect gorillas have my full-hearted support, and perhaps the end justifies the means here in a practical sense, but hermeneutic self-delusion is still self-delusion, whether it’s in the service of a worthy humanitarian goal or a hacker’s fantasy...) SH.]

If one looks at the Schank/Yale work, for example, these NLP systems start with MOVE as a ’primitive’ (i.e. no grounding) and then use this ’primitive’ to build upwards, toward higher-level abstractions, rather than downwards, toward its grounding in the world. Why was this direction taken? Because that was the easiest direction to go (at that time). Now that we can build spike-firing neurons and Fukushima and Crick-style retinas with attention etc, we can begin to explore going "downwards", from the word "move" to its meaning, in terms of that word’s relation to the many experiences of objects moving on the retina. But in doing so, let’s not forget the sensory INVARIANT abstractions of language (e.g. "ownership", "fairness", etc.), for which much (but not all!) of their semantics has to do with their relationships to other abstractions (e.g. OWNERSHIP <-- BUY) than their relationships to sensory experiences.

[SH: My prediction is that if the top-down linguistic approach ever really tries to ground its current systems, the systems will break down hopelessly and everything will have to start all over -- from the bottom up. The work on "MOVE" has been moving up instead of down not only because that's the easy way to go, but because if it moved the other way it would quickly crash. And it's not that the higher abstractions are any different; they only seem to be coherently sustained by their "relationships" because they are safely buffered at their ragged edges by our hermeneutic projections. (There's no sensory infinity there, just a heremeneutic infinity!) Invariant structures there may well be, but they surely won't be these arbitrary, parasitic fragments of word-games, hanging near the top of our lexicon by a skyhook. They will be grounded, all the way to rock (sensory) bottom.]

I am very much aware of BOTH directions, since right now a student of mine (Valeriy Nenov) is teaching a system with a retina to learn the meaning of "move" by observing visual action sequences, and we can see that this task is much easier than teaching the system to learn "owns" (by observing other visual action sequences).

[SH: Because "owns" has a much broader sensory basis and is much higher in the abstraction hierarchy. You don't need a grounded category of "person," for example, to ground "move" (even though we of course know persons can move), but you do need one for "owns." Bottom-up modeling is a tough assignment, especially if the Total Turing Test is your goal. You can't take a short cut by getting ahead of yourself, entering the symbol hierarchy at some arbitrary point in the sky: That just leaves you hanging by a skyhook.

Also, as I suggested in "The Symbol Grounding Problem," the reason for AI's perpetually recurring "brittleness problems" and "frame problems" is probably the unsatisfactoriness of trying to second-guess real-world contingencies and then build the results into the symbolic "structure." The only thing that conceals the fact that top-down, purely symbolic structures are hanging from a skyhook is the persuasive power of the meanings we project onto them: The hermeneutic hall of mirrors. SH.]

So, I'd like to introduce some terminology:

A "symbol" is a pattern that enters into causal relations with other patterns. If the symbol is formed from sensory experiences and (along with various cues) will reconstruct sensory experiences in different networks, then I call this a "g-symbol" (grounded symbol). If a symbol enters into abstract relations with other patterns, that are more logical in nature (.e.g x BUY o --> x OWN o), then I call it an "l-symbol" (logical/relation symbol). If the symbol is capable of doing both, then it's an "f-symbol" (full symbol).

[SH: I don't know what you mean about a g-symbol's being "formed from sensory experiences." In my own approach I am pretty explicit about this: Elementary symbols are the names of object and event categories that the device is able to pick out from the object's or event's projections on its sensory surfaces, using feature-detectors. An elementary "X" would be picked out on the basis of learned or innate sensory features. All higher-order symbols are combinations of grounded elementary ones, e.g.: A "Z" (new symbol) is an "X" (grounded) that is "Y" (grounded).

Now it’s not clear to me whether it’s something like this that you mean by a g-symbol, or merely a symbol "structure" whose interrelations have not just been second-guessed the old top-down way, but supplemented by some induction on actual sensory cases. If it just ends up with a symbolic description -- even if some of its "structure" was induced from sensory input -- I think you’re still left with ungrounded symbol crunching. The grounding must be autonomous and intrinsic. The system itself must be able to pick out the objects and events to which its symbols refer. Our interpretations must not be doing any of the work of grounding the system. Is this the goal of g-symbols? Are they the grounded primitives in this sense? Now I have no problem with symbols for logical relations (conjunction, disjunction, negation, quantifiers, conditionals, category membership and inclusion), but what about the rest of the higher-order symbols? According to me, they must be grounded recursively in the lower-order ones, and ultimately the primitives, by logical combinations (mostly category inclusion). Is that what your f-symbols are? Note, though, that "g-symbols" (in my sense) are just sensory category labels. Their "meaning" is COMPLETELY dependent on their connections with the NONSYMBOLIC mechanisms that pick out the objects to which they refer from their sensory projections. I put meaning in quotations because I too consider it a system property, based on a lot of interrelationships (both sensory and symbolic) that the system only has at the TTT-scale. But I can say right away that if you consider only the "system" consisting of the primitive labels and the higher-order ones composed out of them, you have nothing but a meaningless symbol system again. The meaning is derived at least as much from the nonsymbolic component that picks out the objects as from the abstract and logical relationships among the labels, primitive and higher-order. You cannot sever the connection between the sensory feature detecting system and the symbols or all your left with is meaningless symbols. That’s why I stress that my grounding scheme is nonmodular: It’s hybrid nonsymbolic/symbolic through and through. The primitive symbols are "dedicated" ones, constrained by their connections with (1) the analog re-presentations and transformations of the sensory surfaces and (2) the sensory feature detectors. This is like no ordinary symbol system, in which symbol "shapes" are arbitrary. This is the price you must pay for intrinsic grounding: a nonarbitrary, sensory constraint on the "shape" of the primitive symbols, which is in turn "inherited" by the higher-order symbols: It is this nonsymbolic component that is grounding the system. SH]

Now, dogs and cats have lots of g-symbols, but lack a large number of l-symbols (and lack the ability to process them, e.g. propagating variable bindings). Helen Keller would have a large number of l-symbols but be impoverished wrt to having lots of g-symbols (she’d still have a lot, from kinesthetic experiences).

Here is a question we must address:

2. Does a Turing machine ’simulation’ of Minsky HAVE TO HAVE some g-symbols (no matter how impoverished) in order to pass the standard Turing test?

The answer, I think, is "yes", because we can ask the Minsky ’simulation’ an infinite number of questions of the following sort:

"Imagine you are holding an apple in your hand. How does it feel if you rub it against your cheek?"

"Recall when you splashed water in the tub. You cup your hand more and hit the water faster. How does the sound change? What happens to the water?"

etc.

Without representations that are "sensory" in nature, the Minsky system will not even be able to begin to answer these questions.

[SH: You have given some reasons, I think correct ones, why sensory grounding will be needed to pass the linguistic Turing Test (TT) in the first place. But what you have left out is the fact that it is the nonsymbolic sensory mechanisms of this grounding -- not just the grounded labels riding on top of them -- in which the intrinsic meaning will consist! Cut that off and all you have left is meaningless labels again.]

Here's a (final?) question we must address:

3. Can we have a system pass the "Total Turing Test" but WITHOUT giving the system a bunch of sensory "transducers"? This question is central, because, if the answer is "yes", then your transducer argument is irrelevant to intelligence.

To get the answer "yes" to this question, I must be allowed to interpret "Total Turing Test" as "handling sensory-style experiences" and not necessarily "sensory experiences in OUR world".

Here's how we do it: We construct a SIMULATED ENVIRONMENT, with ITS OWN PHYSICS. We place the simulated program (let's assume, for the sake of concreteness, that it simulates a neural network of some sort) in the simulated environment and now we feed it simulated sensory experiences. This "neural network" must now "recognize" "objects" in its "environment". When the neural network sends signals to its motor neurons, we update the simulated environment so that the organism's motions and sensory experiences are in a causally coherent correspondence to one another.

Our system is now passing a "Total Turing Test", but since the entire world in which the test is being conducted is itself simulated, THERE ARE NO ACTUAL SENSORY TRANSDUCERS INVOLVED!

[SH: Ah, me. You lost the Total Turing Test long ago. You're back into top-down attempts to second-guess nature again. Why not simulate the regular Turing Test too? Don't have real people testing the program. Just second-guess what they'd say and have the simulation test itself, like a chess program playing itself... What will that prove? Absolutely nothing -- except that there's no end to the symbol-games you can play in the light of the hermeneutic hall of mirrors.

Besides, even if, in this grand simulation you SUCCEEDED in second-guessing everything, i.e., even if, as in an airplane simulation, you managed to formalize all the forces of nature and the features of the vehicle so that, using that information alone, you could build a plane that actually flew, the airplane simulation still wouldn't be flying -- and the grand simulation still wouldn't be thinking. An implemented mind is no more a symbol cruncher plus transducers than an aiplane is. Both flying and thinking involve real-world nonsymbolic functions -- staying aloft in the real air in one case, and discriminating, identifying, manipulating and describing real objects in the other. SH.]

Notice, this thought experiment is intended to deal with your (Harnad's) claim that the Total Turing Test requires transducers. It is NOT intended to deal with Searle's argument (i.e. Searle could imagine doing ALL the calculations involved in setting up this simulated world and passing this version of the TTT). (To handle Searle's argument, see first half of this message.)

The physics of this simulated world NEED NOT LOOK ANYTHING LIKE the physics of our world. All that matters is that CAUSAL COHERENCE be maintained (between the neural nets internal representations of this world and how the simulated world reacts to the actions the neural net takes as a result of its internal representations).

[SH: Mike, as I tried to point out in my annotations, Part One did not deal with Searle's argument at all, alas; it just conjured up a lot of Sci Fi fantasies. And now your grand simulation answers neither me nor Searle; it's just a retreat, yet again, to the hermeneutic circle of symbols and our projected interpretations. It has nothing to do with either the letter or the spirit of the TTT. Now I will retire to my tent and let you have the last word...]

So my conclusions are:

1. The "systems reply" is perfectly adequate to handle Searle's chinese box. His argument never dealt any kind of "blow" to the idea that mind is computations.

2. Searle's own introspection is NOT acceptable because, if he (as Searle) can talk about "what it's like to be Searle computing-to- form-Minsky, then there are two entities in one brain, and we will only find out what the Minsky system feels by ASKING the Minsky system (not by asking Searle), and if Searle is creating Minsky "effortlessly and automatically", then in those moments, there is NO Searle; there's just Minsky, since those are the only computations Searle's brain is doing! (There's something uniquely personal about consciousness, see Dennett's discussions on this.)

3. A version of the Total Turing Test (let's call it the SWTTT "Simulated World Total Turing Test") can be passed WITHOUT Transducers. (see the philosophical position of the people doing research on "Artificial Life" in a book with that name, Langton as editor)

4. To pass even the standard Turing Test, some amount of symbol grounding is needed (i.e. to some kind of physics, whether it's the physics of our own world or of a simulated world).

5. A symbol's "intentionality" (for human-level cognition) is a combination of both its grounding to some physics, AND its relationship to other symbols in sensory-invariant ways (i.e. it's I-symbol nature), so it's naive to emphasize just g-symbols over I-symbols, or vice versa, both aspects are needed for human-level intelligence. We need "full symbols".

-- Michael

ON THE NONTRIVIALITY OF TRANSDUCTION (Or, the Boy and the Pony)

From: David Chalmers

(1) The causal intuition (or, "Give me a principled distinction, please")

Clearly, it would be circular to argue "The Chinese Room cannot think, because it is only crunching symbols" and also argue "Symbol manipulation is inadequate, as demonstrated by the Chinese Room" simultaneously, without added argument. One of the two points ("Chinese Room" or "symbol crunching") has to be argued independently, to provide a foundation. In your reply to the Churchlands, you bite this bullet and throw the burden explicitly onto the Chinese Room argument as the cornerstone.

- [SH: No circularity at all: Searle's Chinese Room shows that - symbol-crunching is not understanding, and my symbol grounding - analysis gives some of the reasons why, and suggests a remedy.]

So, if you're going to convince me that the Chinese Room is not thinking, it's going to have to be through other means than an appeal to the limits of symbols. A substantive argument is required, showing that there is a principled distinction between the Room and the brain, showing why one might be "thinking" while the other is not. Such an argument is what I tried to provide for you with "active vs. passive causation."

- [SH: The appeal in the Chinese Room only concerned the absence of - understanding. When Searle has memorized the symbols there's nothing - left in the room to distinguish his brain from. Now, Searle is not - understanding Chinese: That shows that symbol crunching is not - understanding. But he IS understanding english; therefore, whatever - his brain IS doing to accomplish THAT, it's not just symbol - crunching. (I've suggested what else it might be.) These are all the - "principles" you need to understand the argument... The issue is not - "active vs. passive causation" but nonsymbolic vs. symbolic function, - as in symbolic furnaces vs. real furnaces: There's no "passive heat" - in a simulated furnace.]

But instead, you [SH] reply:

> [quote from SH:] Because, (1) as he's stated repeatedly, Searle could
> in principle memorize all that stuff, and then there would be NO
> EXTERNAL APPARATUS at all! Besides, (2) it was never plausible in the
> first place that if Searle alone couldn't understand then /Searle +
> paper-with-symbols/ could understand. "System-as-a-whole" is just
> hand-waving at the ghost that hackers have become all too accustomed to
> freely projecting onto their machines, first just as a mentalistic
> short-hand, by way of analogy with mnemonic variable names in
> higher-level programming languages, but eventually as a habitual and
> uncritical hypostasization that some have even come to regard as an
> "axiom."

OK, so this is your substantive argument. I'll get "(1)" out of the way first, as it's just a chimera. Firstly, it's not talking about the situation in question but another situation entirely, so it's not clear that this argument proves anything (unless you can show that the two cases lie on the same side of some independent metaphysical fence, and thus rise and fall together.)

- [SH: Yes it is talking about the situation in question; in fact, - Searle had already made the memorization argument in his original - 1980 target article, before the years of commentary. And I'm not the - one invoking "metaphysics"; just the simple, common-sense notion that - if a person doesn't understand something by manipulating symbols - on the basis of rules he looks up on a blackboard, he wouldn't - understand it if he memorized the rules either. The friends of the - "Systems Reply," the ones who believe that "Searle + blackboard" - understands even if Searle does not -- they're the ones with the - funny metaphysics...]

Secondly and more importantly, the "memorization" assumption is blatantly counterfactual. Such memorization is far beyond the powers of any human, even "in principle." The required information will have a complexity somewhere around that of the brain itself; to suggest that the brain could double its complexity through memorization is sophistry. (Remembering phone numbers is hard enough; we're talking thousands? millions? of telephone books here.)

If you wanted to, you could cry "thought experiment", and shift the ground to a hypothetical world where brains had a vast "tabula rasa" memory space which was capable of absorbing indefinite amounts of information; such a brain would be so different from our own, however, that our intuitions are worthless. I am comfortable with the idea that the complexity embedded by "memorizing" an entire extra brain's worth could constitute a distinct system with its own distinct phenomenology, and thus "understand" independent of the memorizer. Whether your intuition agrees with this or not, the point is that Searle's vastly counterfactual assumption renders these intuitions worthless.

- [SH: The idea that a person could memorize all those symbols, or even - execute them fast enough, may well be counterfactual, but not - obviously more or less so than the assumption that the Turing Test - could be passed by symbol crunching alone in the first place (for - reasons I've spelled out in "Minds, Machines and Searle" and "The - Symbol Grounding Problem"). So let's not reject some arguments and - not others on the grounds that they might be counterfactual. - - You should also want to avoid being in the position of arguing that - the difference between having and not having a mind is merely a - matter of QUANTITY in some familiar, mindless commodity (speed, - capacity, "complexity" -- pick your poison). That's just hand-waving, - no matter what your intuitions about potential "phenomenology" might - be. And it's hand-waving purely in the service of saving the initial - assumption that "thinking IS just symbol crunching," when that's really - the assumption that's on trial here. There's the real circularity in - this discussion, and I'll even tell you its source: It's the - hermeneutic hall of mirrors I've been mentioning, the one that arises - from projecting your meanings onto symbols in the first place (as - hackers have been doing for too long). Once you let yourself do that, - your intuitions will never allow you to break out of the hermeneutic - circle. - - Here's a remedy: Forget about "thinking" for now; focus your intuitions - instead on some of the indisputably nonsymbolic functions I've - singled out. Remind yourself that symbolic airplanes, furnaces and - transducers, be they ever so turing-equivalent to real ones, cannot - fly, heat or transduce. The ONLY thing that makes you so convinced - that matters are otherwise in the case of thinking is your ASSUMPTION - that thinking is only symbol crunching. Now the only way to - appreciate that simulated airplanes, furnaces and transducers don't - really fly, heat or transduce is by stripping them of their - interpretations and recalling that, after all, they are really only - crunching symbols in a way that is INTERPRETABLE as flying, heating - and transducing. The uninterpreted system is clearly doing nothing of - the sort. This suggests that the right way to examine whether a - symbol system is really thinking is to consider only the - UNINTERPRETED system, without the hermeneutic help, and see whether - there's any thinking going on there. This is exactly what Searle has - done. And the answer is: No. The metaphysics about "systems" is just - a desperate attempt to resurrect the original assumption that this - outcome has knocked down. (Try the "Systems Reply" on simulated - flying, heating and transduction...) - - By the way, some of my arguments for the advantages of nonsymbolic - (e.g., analog) representations turn the Church-Turing Thesis on its - head in a way that is analogous to the capacity/complexity - counteragument YOU are making here: Though simulable symbolically in - principle, a lot of the analog sensory functions of the brain may - require a much bigger brain to house their turing-equivalents - instead. Moreover, these turing-equivalents would clearly also have to - include a simulation of all the possible contingencies the outside - world could have visited on the original analog functions they are - replacing. A lot to fit in one brain. - - And of course the transduction done by the sensory surfaces - themselves cannot be replaced by simulation even in principle - (unless we are contemplating simulating the entire outside world - too -- which would then make the whole exercise homologous and - not just analogous to simulated flying, even by definition!). - - So even your own capacity considerations point toward the conclusion - that a thinking brain (or its TTT-equivalent) is no more likely to be - only a symbol-cruncher plus transducer surfaces than a real plane or - furnace is. Transduction, in other words, may not be trivial; in fact, - a lot of the function underlying cognition may BE transduction. SH.]

Moving on from this distracting point, your point "(2)" suggests "it was never plausible in the first place that if Searle could not understand, then Searle + paper + rules could." Now this is the real meat, and this is where I'd like to see some argument. Instead, all I see is the usual casual dismissal of symbols and the "Systems Reply." More is required here (though I've never seen it in either your writings or Searle's). We have two systems to compare: the human brain and the Chinese Room. Both are complex systems, which undergo various processes over time, leading to intelligent-seeming behaviour. One of these certainly understands. You still need to provide a *principled distinction* demonstrating why the other does not.

- [Ah me. The logic, once again, is this: Brains can pass the TT and the - TTT. Let us suppose (possibly counterfactually) that a - symbol-cruncher could pass the TT. Is it doing what the brain is - doing? Is it understanding? With the computer sitting out there, the - question is moot, because of the other-minds problem. The only one - who can know whether or not the computer is really understanding is - the computer. So Searle says: let ME be the computer. Let me do - everything it does, and let's see whether I understand. He doesn't. - - What do those who believed that symbol-crunching IS understanding - reply? Searle doesn't understand, but "Searle + rule-book" does. A - curious gambit, because until now the other-minds problem at least - had a locus. You could say who or what was the entity that might or - might not be doing the understanding, its subject, so to speak. But - Searle plus rule-book? Distributed understanding? - - In any case, Searle says, fine, I'll not only DO what the computer - does, I'll BE what the computer is by memorizing and hence - internalizing the rule-book. So now there's nowhere to point to but - me, and I still don't understand. - - What's your reply to this? "You couldn't memorize the rule book. It's - too big. And if you had the kind of brain that COULD memorize the - rule book, you WOULD understand." (Others, trying to save the - Systems Reply -- which is really just the original "thinking is just - symbol crunching" assumption, repeated more loudly this time -- reply - that memorizing the rules would induce a multiple personality in - Searle even with a regular-sized brain; see Mike Dyer's replies; - still others say the rules themselves undetrsand! Sci FI in the - service of desperation knows no bounds.) - - I just want to point out that Searle has had to use no Sci Fi at - all; only simple, straightforward notions like executing or - memorizing symbol manipulation rules -- exactly what a symbol - cruncher does. All the Sci Fi is coming from the other side, in an - effort to save the original intuition (that if it does all these - smart things that are interpretable in a meaningful way by us, then - the symbol cruncher must have meaning and understanding too). A more - reasonable, nay, LOGICAL response would seem to be either to concede - that the interpretation turned out to be unwarranted, on the - evidence, or to come up with counterevidence. Sci Fi is not - counterevidence. It's just pushing off further and further into the - hermeneutic heights. SH]

As I’ve said, an appeal to "symbol crunching" will be question-begging. You’ve rejected the appeal to "active vs. passive causation" (though I claim that this is all you can mean by "The human is doing all the work"). Searle sometimes appears to make an appeal to "biology", but does not back it up. Certainly the "Systems Reply" is not a knock-down argument, but the onus is on you to show just what’s wrong with it, and to point out the distinction between the systems that might make such a vast metaphysical difference.

- [SH: I have done just that, although this may not be visible until you - step out of the hermeneutic hall of mirrors: The evidence shows - that understanding is NOT just symbol crunching, despite the fact that it - is INTERPRETABLE as such. And the crucial distinction has to do with - groundedness, which includes transduction, a nonsymbolic function. - Searle has that (for English), whereas the Chinese symbol cruncher, - whether implemented by the computer or by Searle, does not. They only - have it projected on them by you. - - In the hermeneutic hall of mirrors, though, I’ve answered the wrong - question, because the question is not what function OTHER than symbol - crunching might understanding be -- of course it’s symbol crunching, - what nonsense to supppose otherwise -- the question is only "Where’s - the understanding? If it’s not in Searle, it’s gotta be around here - someplace...." The situation reminds of nothing more than the boy - looking in the proverbial pile for the pony...]

The rest of your reply here consists of an appeal to the inadequacy of symbols. This just leads to a circular argument, so I’ll disregard it. (The other option open to you is to base your argument *completely* on the inadequacy of symbols. This would be fine, but then the Chinese Room would be completely irrelevant.)

- [SH: As we know from Achilles and the Tortoise, you can’t prevent - someone from ignoring anything he’s intent on ignoring. There’s no - circularity here whatsoever, apart from the hermeneutic circle you - seem to be trapped in.]

(2) The semantic intuition (or, "Which symbols shall we ground?")

I won’t go any further with this one, as it’s the kind of discussion that’s likely to go round and round without making progress. I’ll just note that if for you, "symbol" means "formally manipulated object", then connectionist symbols have no need to be "grounded", any more than neurons do. No meaning was ever intended to be present at that level. (And thus "Symbol Grounding" is the wrong term.) Neurons aren’t grounded either. The brain thinks, despite the "Neuron Grounding" problem.

- [SH: You’re again just repeating your assumptions. There’s no neuron - grounding problem only because brains are not just doing symbol - crunching (the assumption you can’t let go of). And as I wrote in - "The Symbol Grounding Problem," I’m not using the word "symbol" - figuratively. There are eight conditions symbols must meet in order - to be symbols in the formal Church-Turing-Newell-Fodor-Pylyshyn - sense, and this includes compositeness and systematicity. If - connectionist "symbols" don’t have those, then they are not symbols. - They are merely interpretable in isolation as "symbols." (Anything - at all can be interpreted in isolation as meaning anything.) - - But whether or not nets have formal symbols is not really relevant; - if nets can pass the Chinese TT, the question is only whether or not - they are susceptible to Searle’s argument. Simulated on a symbol - cruncher (as virtually all current nets are), they certainly are - susceptible to it. Now, whether there is something ESSENTIALLY different - for mind-supporting purposes in a parallel implementation of the same - net (which Searle of course cannot duplicate, just as he can’t - duplicate transduction) calls for some argument, it seems to me, as - to why the implementational difference matters even though there is - I/O equivalence.

- For in the case of flying, heating and transduction, there is no I/O - equivalence between their respective symbolic simulations and their - real-world implementations, just computational equivalence (turing - equivalence), because flying, heating and transducing draw upon - essentially nonsymbolic functions. Now parallelness too is - nonsymbolic, but the I/O function directly at issue happens to be - TT-power, not parallelness. It's obvious why transduction is - essential for TTT-power, but why should parallelness be essential for - having a mind, on the strength of the TT, if the TT could be passed - by a symbolic simulation of the parallel net too? (Note that I'm not - saying the difference could not be essential, just that there's some - burden to show how and why in the case of parallelism, whereas - transduction wears its functional reasons on its sleeve.) SH]

(3) The Turing Test (or, "Accept the consequences")

> SH [prior quote]: Please follow this carefully; it will prevent further
> misunderstandings: First, I have never ARGUED that a pure symbol
> cruncher could not pass the TT in principle., i.e., that this is
> logically impossible.

Wonderful. You buy Church's thesis. I wouldn't have guessed. Just to reinforce it, repeat after me "A symbolic simulation could pass the Turing Test." No ifs and buts, no hedges.

- [SH: Sorry, the best I can do is: I know no proof that it would be - logically impossible for a symbol cruncher to pass the TT; hence the - TT might in principle be passed by a symbol cruncher. I also know no - proof that it would be impossible for a person to memorize its rules - and symbols. But I think it highly unlikely that either could be done - in practice. - - Church's thesis, by the way, is NOT the thesis that thinking is symbol - crunching; just that every "effective procedure," including every - physically realizable process, is simulable by a symbol cruncher. Nor - should this formal "turing equivalence" be confused with the Turing - Test, which is just an informal, open-ended test of our intuitions - about other minds under conditions of I/O equivalence - (symbols-in/symbols-out in the TT, and all of our robotic and - linguistic I/O in the real world in the TTT). And Church's thesis - certainly does not entail that every physically realizable process IS - symbol crunching: There's still flying, heating, transducing, and, - despite the ambiguity of the TT and the lure of hermeneutics, - thinking.)]

After making this admission, you go on to hedge it with all kinds of vague appeals. For instance:

> [SH:] successfully passing the TT would probably have to draw on the
> functional capabilties needed to pass the TTT, and hence that whatever
> could pass the TT would also be able to pass the TTT -- and THAT, as I
> have shown, CANNOT be all symbolic, because of the sensory projection
> at the very least (transducer function), and probably a lot more like
> it too.

I don't know exactly what you're trying to say here. A symbolic simulation, you've conceded, *can* pass the TT. Maybe it does it through symbolic simulation of "non-symbolic" function -- I don't really care. The program itself is symbolic. Now, if you say "anything that can pass the TT can pass the TTT", then it looks like symbolic simulations are home free.

- [SH: No, you’ve gotten the logic wrong. My ambiguous wording is at - fault. I should have said "...hence that whatever could pass the TT - would also already have to have been able to pass the TTT..." The - point is very simple: We have supposed that a pure symbol cruncher could - pass the TT (because of Church’s thesis, say), but then Searle shows - that, if it could, it would not be understanding. So we scratch our - head and ask why a pure symbol cruncher would not be understanding - even if it could pass the TT. Maybe it has something to do with - grounding the symbols in the world of objects they stand for. - - So we turn to the robot who can pass the TTT (the test that we know - is immune to Searle’s argument) and who -- by my lights, if not - Searle’s -- really DOES understand, and we say: Maybe some of the - nonsymbolic functions of that understanding robot are drawn upon when - WE (and those like us, who can likewise pass the TTT as well as the - TT, and likewise really understand) are passing the TT. In other - words, even if a symbol cruncher alone could pass the TT, it wouldn’t - understand, because of Searle. Yet when we pass the TT we do - understand; perhaps this is because we, unlike the pure symbol - cruncher, are drawing on our nonsymbolic TTT-functions in order to pass - the TT. (And if the premise that a symbol cruncher could pass the TT - is counterfactual after all, maybe this is the reason why!) - - Now although, again by Church’s thesis, the TTT-passing robot could - be simulated by a symbol cruncher, the simulation could not itself - pass the TTT, any more than a simulated airplane could fly or a - simulated furnace heat. (And if it passed only the TT, it would still - be just another symbol cruncher passing the TT -- mindlessly, as - Searle has shown.) At the very least, to pass the TTT the robot - simulation’s simulated transducers would have to be turned in for - real transducers (and its simulated world for the real world), which - means that the system would no longer be a pure symbol cruncher. But - I think the analogy with flying and heating is deeper than that. And - the answer to the argument that "transduction-is-trivial: even a - computer needs transducers, so just hook ’em on up!" (which you are - no doubt waiting breathlessly to make) may be that implemented - thinking is no more a matter of hooking a symbol-cruncher to - transducers than implemented flying or heating is. SH]

After making the concession, you further proceed to back away from it with "I don’t know if a symbolic simulation really could reproduce the essential properties of non-symbolic function." I’m sorry, there’s no "I don’t know" about it. If you buy Church’s Thesis, then they *can*. We’re talking about complete simulations here, not "kind-of" simulations. (When you say "a sensory analog may be worth more than an infinity of symbols", this can be interpreted as nothing but a rejection of Church’s Thesis. Such a strong claim sounds simply naive. Many physicists have thought about these issues in great depth, and there’s little evidence supporting you.) Anyway, state a position: either physical processes can have perfect functional simulations or they cannot. If "yes", then your hedges are irrelevant. If "no", then you’re on very weak ground.

- [SH: No hedges: Perfect functional simulations they can all have -- - planes, furnaces, transducers, thinkers -- but not real-life - implementations that actually fly, heat, transduce and think, - because not all physical functions are IMPLEMENTABLE AS symbol - crunching even if they are all SIMULABLE BY symbol crunching. - No physicist, be he ever so deep, can deny that. (The analog issue is - just the capacity problem discussed earlier in this message; I won’t - force the issue if you don’t.)]

Steve, you’re on stronger ground making metaphysical claims than operational ones. I’d stick to metaphysics. The operational claims come out meaning precisely nothing.

- [SH: No, I don't need the metaphysics; besides, to me what you're - calling "metaphysics" looks a lot more like arbitrary Sci-Fi - stipulations to save a hermeneutic assumption -- the assumption that, - because symbol crunching is INTERPRETABLE as thinking, it must BE - thinking.]

(4) The Total Turing Test.

Very briefly, a thought experiment. A robot reproducing the external details of a human being, equipped with appropriate I/O transducers all over the body. Input from these transducers are used as "boundary conditions" for a numerical system of differential equations which are meant to model the physical processes in the human body. (Church's Thesis implies that such a model is possible.) It doesn't matter whether, in the body, such processes are "nonsymbolic"; we'll simulate them symbolically. As the system evolves, we'll have simulations of output signals going to the periphery of the body, which can then "transduced" into real output actions on our robot. Thus, we have a perfect *TTT*-passer, which consists of a sensory periphery with a symbolic core. Possible, that's all.

- [SH: Agreed that it sounds possible, though far-fetched. Now, what - Searle's Chinese Room Argument and my Transducer Counterargument have - shown is that, if such a system is possible and understands (and I - do happen to believe that if it is possible then it does understand), - then the functions in which the understanding consists -- the - physical processes that have to be going on for understanding to be - going on, and the functions you have to BE in order to be - understanding -- cannot be just the symbol crunching: They must be - the symbol-crunching AND the transduction. I simply think (for - reasons given in both of the papers under discussion) that in reality - the nonsymbolic/symbolic ratio in a TTT-passer (including ourselves) - will have to be be the exact reverse of what you've just conjectured, - and that transduction will turn out to be anything but trivial. - - However, I am perfectly satisfied with the logic even in your - counterfactual version, for even there, understanding is still not - just symbol crunching. (Now YOU repeat after me...) SH.]

"Non-symbolic all the way in" means nothing. We can simulate it fine. The *only* point at which analog is essential is at I/O. Arguments to the contrary are unfounded. (It's true that to get a good enough simulation we may have to go to a rather low level; but that's the point which is beginning to be made by connectionism, after all.) -- David Chalmers.

- [SH: You can simulate it all (except the transduction) in principle, - but what about in practice, with real TTT-scale robots? And what if - being in the state of understanding draws on internal functions that - are like heating or flying -- or even essential parallelism? I - personally don't care, as long as the robot can pass the TTT, though; - so I'll settle for your concession that symbol crunching alone could - not be thinking: that at the very least, analog TTT I/O functions are - essential to implement thinking. SH]

From: Dean Z. Douthat Subject: Motor learning

I have followed with fascination your ongoing debate with the symbolic bigots and must say I admire your patience in the face of persistent refusal to recognize and truly respond to your arguments. As the article I sent you on the "Life of Meaning" shows, this basic argument extends to very low levels and covers extremely broad grounds. The essential differences between analog systems and digital systems, including digital systems simulating analog systems, seem to be major stumbling blocks for people whose education and experience have not included actual contact with analog systems. Perhaps there is an educational issue here meriting attention from computer science programs. Perhaps it was a mistake 15 or 20 years ago to eliminate analog computing as part of the CS core.

So ingrained is the symbol and/or number crunching mind-set that ridiculous extremes will be accepted to defend it. Besides the transduction or sensory forms of analog processes which you take to be essential to thinking, I would also advance motor "learning" as equally important at the other end. And the interactions between them may be, but are not necessarily, symbolic. The example I give is that of a major league centerfielder turning at the crack of the bat, running back to the wall and pulling in the fly ball over his shoulder. Believe it or not, I have met many symbolists who believe his brain is actually solving a set of differential equations to arrive at a trajectory. Then he simply runs to where the trajectory intersects the plane parallel to the ground at say shoulder height. Thus does Willie Mays become the greatest mathematician alive.

Of course his actions can be INTERPRETED as if he solved equations but that clearly is not happening. For one thing, he has insufficient initial value data even to come close to such an approach. All your arguments against the "systems reply" apply here as well. Perhaps this example has the advantage of being less directly tied to emotionally held preconceptions than is thinking, in that it does not open the possibility of non-material, mentalist entities. It also has the advantage that if a symbolic bigot persists, the example can be converted to a dog catching Frisbees, to antelopes performing optimum grazing [i.e. solving dynamic programming problems] to bacteria finding best available chemical surrounds, etc. As the argument is driven to lower and lower forms of life, symbolists abandon ship one by one. Eventually none are left.

Dean Z. Douthat

- [SH: I like your point about motor learning, and although I have - mostly spoken about transducers in this discussion, in the published - versions I have always stressed transducers AND effectors. I expect - that the motor end is at least as nonsymbolic as the sensory end, - and that there's very little room for pure symbol crunching in - between. (I have never alluded to "non-material, mentalist - entities," by the way; it is only that symbolists have somehow - convinced themselves that to deny that mental processes are - symbol manipulation is tantamount to advocating Voodoo! No such - thing; nonsymbolic functions are just as monistic and physical as - symbolic symbol manipulation.) -- Stevan Harnad]

From: miken@ai.mit.edu (Michael N. Nitabach) Re: Dean Z. Douthat Subject: Motor Grounding

As my day to day work is on computational methods in motor control I was intrigued by Dean Douthat's remarks. I wholeheartedly disagree with his interpretation of the computations used by a baseball player in catching a fly ball. I challenge Dean to give me *any* alternative explanation of the performance of this task that does not involve the computation of some approximation (perhaps greatly simplified) of the equations of motion of a ball in flight. His remarks depend on the intuition that one does not consciously experience oneself calculating trajectories when playing baseball. This is a useless and dangerous type of intuition to apply to the analysis of cognitive capacities, I hope Dean is not claiming that the only *real* cognitive processes are those that are accessible to consciousness; unfortunately, that seems to be the tacit assumption he makes. In short, Dean has simply confused the distinction between "tacit" and "explicit" knowledge, with the distinction between "real" cognitive processes and those that are only "interpreted" as occurring. He claims that the illusory nature of the supposed computations is "clear." He expresses disbelief that anyone could think that a baseball player *is*, in fact, performing computations over a model of ballistic flight, albeit a simplified one.

This is evidence of Dean's lack of contact with the actual work going on in the field of motor control at the present time, e.g. in the learning and planning of arm movement trajectories, and locomotion. As I said before, I challenge him to support his very strong dismissal of a huge body of experimental and theoretical work on motor control (and "physical intuition"'s cognitive basis) with either (1) an analysis of the actual content of the field and arguments against the prevalent view and/or (2) some real argument in support of his own view, not just appeals to intuitions about cognitive processes that have been shown time and time again to be false.

I would be happy to enter into a discussion of the relevance of various types of evidence for the attribution of reality to hypothesized cognitive processes, provided also that Dean slightly lessens his scornful attitude.

Michael Nitabach

- [SH: I am prepared to mediate this discussion between Mike Nitabach - and Dean Douthat (and between Mike Dyer and Dean Douthat, see below), - which has materialized as a spinoff of the Symbol Grounding - Discussion. However, I agree with Mike that as a precondition all parties - must refrain from using expressions like "symbolic bigot." - Dean: as you know, it appeared more than once in your original message, - but in posting it to the discussion group, I left in only the first - mention, because there it could perhaps be construed as a good-natured - joke. It evidentally has not been so construed, so, please: only - gentle, non-ad-hominem contributions henceforth. SH.] -
----

From: Dr Michael G Dyer To: Dean Z. Douthat Cc: harnad@clarity Subject: "symbolic bigots" and motor transducers

The term "symbolic bigots" has an unfortunate ring of nastiness about it. I hope that those involved in this discussion will at least pretend that they respect those who hold opinions differing from their own.

Modern computers are made of a whole bunch of transducers, which both receive forms of energy as input and produce forms of energy as output. These transducers may be viewed as analog OR digital, depending on at what level of behavior you want to examine their physical properties, and depending on whether you want to view the fundamental equations (describing the physics of the universe) as digital/quantum or analog/continuous. I've talked to a number of physicists about this analog/continuous vs digital/quantum business and they emphasize that it's not what reality is really like, but how one is choosing to DESCRIBE reality and one can pick discrete or analog languages and they're roughly equivalent.

Now, it turns out that when people build digital computers, they put these transducers together in ways to make the components act in discrete manners (i.e. they ignore the analog properties). Supposedly, any analog system has a precision of some sort and it can be simulated (sometimes only with tremendous effort) by a digital system.

As far as I can tell, the issue of transducers is not that they are "analog" but that they have physical effects on the world (motor) and pick up energies from the world (sensory) that the transducers used in computers cannot do.

From my "computational systems" point of view, the transducers are NOT essential for understanding the essence of "seeing". If I can build a simulated retina of spike-firing neurons (on a computer) and set up firing patterns that are similar to what some real retina would do, then I can ignore the real retina, build my computational system to process the patterns coming in from the artificial retina, until the modified patterns go to something to drive a nexus of motors (on output) that are something like what real muscles would receive as signals.

If I can get such a system to work in an appropriate way (e.g. duck an object thrown at it, etc.) in the simulation, then I should be able to hook the system up to a real retina and real muscles and have a real robot duck a real object thrown at it. But, while the real robot is the final test, I can figure out all the essentials of vision within the simulation (assuming that the input patterns on the artificial retina and motors are close enough to what real retinas and muscles are doing as transducers).

So I view myself as a "computational systems" person (since I'm not sure what a "symbolic bigot" is), by which I mean that I view intelligence, consciousness, intentionality, life and associated subjective qualia as emerging as complex systems phenomena, where the essential aspects of the nature of Life and Mind can be examined in terms of computations (numeric/symbolic). Clearly, one needs some physical "stuff" with which to realize these computations, but the particular nature of the stuff is not important, as long as it supports the computations needed. (so we can have a computer that's optical, electrical, chemical, a hybrid of these, etc.)

Some people who are founding the new field of "Artificial Life" (AL) take the same position, but for Life itself -- that life (and all its complex biochemistry, genetics, populations and environments, etc.) can be modeled as computations. Other AL-ers take bascially the same position that Searle takes for AI -- i.e. that behavior the is "life like" does not mean it is actually alive.

An extreme position within the "computational systems" perspective would be that essential issues of Matter can also be described in terms of computations -- that at the "bottom" of reality (smallest unit of space/time) there is just information and the entire universe is a big "game of life"-style machine; i.e. that information is more fundamental than energy. E.g., Fredkin at MIT takes this position quite seriously.

If many people find the claim that computations in devices (other than brains) can someday "be" (not just act) intelligent, then how much more shocking the claim that computations in devices (other than natural populations of cells) can someday "be" alive (and not just act as if alive).

Within this debate there is a AL-Searle. The position of an AL-Searle will be that, no matter how many features of life such simulations exhibit, they won't "really" be alive, since they are not of the same "stuff" of which life is made (E.g. the "waste" products of the AL creatures won't smell the way real fecal matter smells -- but what if the simulation includes simulated chemical molecules and simulated smell by simulated neural networks and simulated olefactory sensors ??)

When Harnad states that the patterns the organism manipulates internally need to have a causal correspondence with patterns being received by input transducers, then I agree. When Harnad states that without actual transducers the system has "only meaningless symbols" I then disagree, since I can bypass the real transducers in favor of simulated transducers that act like the real ones (with respect to what they send to the rest of the system). "Grounding symbols" to me means having a correspondence between the information processing system and an environment, but the environment can be an artificial one.

You know, symbols, while we view them as abstractions, actually have physical realizates, in the form of patterns of energy/matter. That's why Newell talked about "PHYSICAL symbol systems". We just talk about them as "symbols" because we're usually interested in their configurations (patterns) and not their physical embodiments (since there are many ways of embodying them).

-- Michael Dyer

- [SH: I find that some of this is right, but much of it is wrong - (particularly the stuff about pure simulation), but Mike and I - have already said our piece to one another. Now I'll leave it to - Dean to see what kind of an account he can give of himself. SH]

Hi Eric, I read the Computers & Philosophy 1988 article. The argument was the same as in the Computationalism paper -- that computation is not just symbol crunching, but depends on the meaning of the symbols, and that symbol crunching is meaningful because in order to understand it we must interpret it as meaningful.

I think I've refuted these points in my commentary, Computational Hermeneutics. I'll be interested to see your response. The gist of my argument is that variable lookup is quintessentially syntactic (symbol manipulation based on symbol shape alone), not semantic. It only looks semantic if you project meaning onto it. And of course you can only understand formal systems if you project meaning onto them, i.e., if you interpret them, but so what? It is already part of the DEFINITION of a formal symbol system that it must be semantically interpretable. We don't bother with "symbol systems" that aren't.

The bottom line is that symbol systems, be they ever so semantically interpretable BY US, do not contain any semantics: Their only constraint is syntactic. They have no content at all, except what we project onto them. Hence they cannot be the model for our thought processes, because those thought processes clearly do contain their own semantics. I don't mean "bird" merely because I am interpretable as meaning bird. I really mean bird. And that's the real problem of modeling mental processes: Their meaning must be intrinsic to them, not parasitic on someone else's mental processes and the interpretations they project.

You keep replying that computation has meaning because it's interpretable, and would be incomprehensible unless we interpreted it. That simply begs the question of how symbol meaning is to be grounded within the symbol system itself. It would be helpful in your response if you acknowledged explicitly that THIS is what is at issue for those who point out that symbol systems have no intrinsic meaning, and then confronted the problem directly. Otherwise you will just keep bypassing the objection (and missing the point).

Cheers, Stevan

ON COUNTERING COUNTERFACTUALS

David Chalmers wrote:

Steve, this won't be point-by-point, otherwise the discussion would explode and never converge.

On symbol-crunching:

Close examination of your arguments reveals precisely two arguments that might be called substantive:

(1) The "memorization" argument. (2) The "simulated planes don't fly" argument.

Briefly, re (1): when Searle's opponents reply with the "two minds in the one brain" argument, you accuse them of "Sci-Fi". But it was Searle who introduced the Sci-Fi. A world of beings who are capable of internalizing by memorization the entire complexity of an extra human brain is as much a science-fiction world as one populated by little green men. We should expect Sci-Fi assumptions to lead to Sci-Fi conclusions. As I said, I'm perfectly happy with the idea that that vast added "memory capsule" might support an extra mind.

- [SH: Sorry, but that's not where the Sci-Fi started. It started with - the assumption that a pure symbol cruncher could pass the Turing Test - (i.e., talk back and forth for a lifetime exactly as a real person - would) in the first place. Once that's accepted (on the strength of - the Church/Turing thesis, say), then the rest of the Sci-Fi ON THIS - SIDE OF THE FENCE is really quite minimal, only having to do with - matters of speed and capacity. The double-mind riposte, on the other - hand, is launching off into a new dimension of fantasy altogether.

- But never mind: parity's parity. The Sci-Fi began with the TT - assumption, not downstream. One could dismiss the entire issue by - denying that the TT could be passed by symbol crunching alone, so the - whole thing is counterfactual: Do you want to do that? If not, you - can't fault Searle for his initial piece of Sci-Fi (which he in any - case didn't invent, but carried over, arguendo, from Strong AI, to - show that they were wrong even on their own terms).

- Now, parity's parity, so this should be enough to answer your - objection. But since we seem to be niggling about the quality of - counterfactuals, do you perhaps notice any difference between the - following two conjectures that goes a bit beyond mere quantity:

- (1) The quantitative conjecture that, even though present-day - toy symbol systems (chess players, psychotherapists, scene describers, - text paraphrasers, etc.) cannot pass the TT, MORE of the same -- - more speed, more capacity, more complexity, more code -- will - eventually succeed.

- (2) The qualitative conjecture that if a symbol system succeeded in - (1), it would have a mind, just like you and me. (And, if (2), then, - (2c): the corollary that if a person memorized the symbols, he'd have - TWO minds.)

- It's (2) that was tested by Searle (and failed). Then (2c) was - desperately invoked from its ashes to resurrect (2). That's what I'm - calling Sci-Fi. And it simply amounts to circular theory-saving even - in a counterfactual argument. SH.]

(If you come back with the argument, I reply: a mind. Three neurons don't support a mind; 100 billion do. Although complexity is not a *sufficient* condition for a mind, it is certainly a necessary one.)

- [SH: As I've said before, "more neurons" is not the same as "more - symbols" (nor is the difference between a sea slug's brain and a - human's simply that the latter has more neurons...). Nor do we know - what neurons actually do in the mental economy, nor that they are the - relevant functional unit in the brain, nor that a 3-neuron organism - has no mind; we don't know much of anything about neural function, - actually. So we don't even know what we are affirming or denying when - we talk about "more" of THAT. But, thanks to Searle, we at least know - that, since symbol crunching is not sufficient to support a mind, - whatever the relevant neural "more" might be, it won't be just - symbol crunching!

- But if you want to force the issue, fine: I don't believe that a mind - is simply a matter of neuronal quantity, any more than a furnace is - just a matter of molecular quantity. The functional story IS more - complicated; and that complexity is not just symbolic in either case. - (Please don't come back with the tedious story about molecular - thermodynamics: The requisite complexity is still not symbolic.) - - Finally: we are indeed looking for sufficient conditions for having - a mind, not just necessary ones. Symbol crunching was suppposed to - be sufficient for understanding, not just necessary, remember? SH.]

There is a simple, easy to articulate reason why simulated planes don't fly, there is a *principled physical distinction* between things which fly and things which don't fly, roughly "something which flies must undertake controlled motion through the air". A *physical* distinction, note, that planes fall on one side of and simulated planes on the other. Similarly, to heat, an object must "raise the temperature of its surrounds". Another simple, physical criterion.

If you could give me such a simple physical criterion to distinguish thinkers from non-thinkers, and which consequently distinguished people from non-thinkers, I'd be impressed. But I don't think you could do it, because the criteria for "thinking" aren't physical, they're functional (even "logical", if you like). The best physical criteria you could come up with would be behavioural, and I'm sure you don't want to do that.

So there's no analogy between simulated planes and simulated minds. It's obvious why simulated planes don't fly: there's a simple physical criterion which they fail. There's no such criterion for minds (except things like "understanding", and then we're back to the first argument again).

- [SH: To give you the "criterion" you request would be to give you a - true theory of how physical systems manage to have minds. A tall - order (to ask that it be "simple" is even taller). I’m working on it. - But in the meanwhile, may I point out a little problem with your - logic? The fact that we don’t yet know what the critical physical - "criterion" is that distinguishes systems that have minds from those - that do not is not very strong grounds for assuming there isn’t one, - and pre-emptively concluding that therefore it’s all symbol - crunching!

- The Total Turing Test (TTT), which is a "criterion" I advocate, is of - course behavioral. It is because of the other-minds problem -- which - is unique to mental processes like thinking or understanding, and - does not afflict physical processes like flying or heating -- that - the TTT cannot ever be decisive. For whereas anyone can know (as - surely as one can know anything about the world) that a plane is - flying or a furnace is heating, in the case of another mind, the only - way to know that it is thinking or understanding is to BE it (which - is a taller order than any empirical theory can fill).

- Finally, as I said before, the TT is a criterion according to which - Searle has shown that symbol crunching FAILS to produce a mind, - whereas the TTT is a criterion that I have shown pure symbol crunching - could not meet even in principle because, for example, transduction - is nonsymbolic.

- So what have we learned? That (i) understanding can’t be just symbol - crunching (Searle), and that (ii) symbol crunching alone can’t pass the - TTT. So if, for example, transducer function is one of those fateful - physical processes that distinguish thinkers from non-thinkers, then - we already have the "criterion" you asked for! It needn’t stop there, - though. There might be a lot of other nonsymbolic processes that a - thinker must BE in order to be thinking. Give it time and we’ll find - them for you. For now just be content with the interim result that - it’s not just symbol crunching. SH.]

On Tests, Turing and otherwise:

Your operational claims don’t look in very good shape either. By appealing to the Church-Turing thesis, I’ve established (and have heard no substantive defence) that

(1) A symbol-cruncher could pass the Turing Test. (2) A symbol-cruncher with a sensory/motor periphery could pass the Total Turing Test.

I see no content left in your operational claims, then. If you still want to hold onto the TTT as a test of mentality, then (2) above forces you to agree that the Robot Reply is a perfectly good answer to Searle.

- [SH: No such thing. We know that a pure symbol cruncher that passes - the TT has no mind. We know that a pure symbol cruncher without - transducers cannot pass the TTT. Hence it looks as if transducer - function is part of mental function. The conventional robot reply - (not to be confused with my own "robotic functionalist reply") was - just that you had to ADD ON transducers (to pass the TTT). My point - is that, since you have no mind without them, and you have a mind - with them, it looks as if you have to BE your transducers in order to - have a mind. I’ve also given plenty of reasons why this is not just - an "add-on" matter, and that there are probably plenty of other - nonsymbolic functions a mind has to be. But this much is already good - enough to put pure symbol crunching out of contention.]

I might now even, just for dessert, say: take our TTT-passer as above, "blind" it and paralyze it. All we've done is cut off I/O at the periphery (remember, I established that analog is *only* required at the periphery). Surely it's still thinking (if you wanted to, you could maintain that speech, sight, and motor movement are essential, but it would be implausible. I still think when blinded and paralyzed). But it's now effectively a pure symbol-cruncher. Oh well. -- Dave Chalmers.

- [SH: Ah me. Here we are, freely commuting between neurons and - symbols again. In the case of the real brain, to blind and - paralyze it is not to strip transducers off a symbol cruncher. - So what happens to YOU when you're blinded or paralyzed is - completely irrelevant. Most of what you've left intact is just - analog re-projections of the sensory surfaces. If you really peeled - off all the sensory and motor tissue, you'd have very little brain - left. Most of the brain is just ramified transducer/effectors.

- But never mind that; let's go back to Church/Turing-land: Let me - speak in the language of your fantasy: It has now been shown that - /symbol-cruncher-alone/ has no mind, whereas - /symbol-cruncher-plus-transducers/ does. It follows quite - trivially that /symbol-crunchers-minus-transducers/ again - doesn't. QED. - - Dave, I think we've reached the point where we've us shown one - another our conceptual wares, and little new is emerging. We won't - convince one another, so we can now step back and let our - respective arguments be weighed by others. - Cheers, Stevan Harnad]

To: pawlicki@Kodak.COM Subject: Re: understanding

Ted, you wrote

> Are you in fact saying that there is no understanding of arithmetic in the
> previous example? (I'm sorry to repeat myself, but I would like a straight
> answer). If there is no understanding, then how do the correct answers come
> out? Are you saying that there is something more to understanding aritmetic
> other than knowing the symbols and the rules which relate the symbols to each
> other? What is that something?

OF COURSE there is something more to understanding arithmetic than being able to manipulate symbols according to rules! Since when was a desk-calculator a psychological model? The correct answers come out because the rules are followed, and the rules work. That's what a (successful and computable) formal symbol system is: A mechanical procedure for generating symbols that are systematically intrepretable as "correct answers." Now, that interpretation itself is NOT PART OF THE FORMAL SYSTEM, whereas it is most certainly an essential part of our understanding.

As to what "that something" is in which our understanding consists, over and above symbol manipulation: that's a job for substantive psychological theory. I've made my own bid, with my hybrid model for picking out the objects the symbols refer to; but all you need in order to understand the point at issue (and if by now you don't, I'm afraid I can't continue with these exchanges, because they are making no progress) is the first two sentences in the preceding paragraph.

If you are already so committed to the notion that a desk calculator understands arithmetic (or anything) that you haven't noticed that this is the very notion that has been under fire from counterexamples and counterarguments and logical analyis in the symbol grounding discussion, and that simply reiterating it is not an answer -- then this exchange is really pointless. You may as well also tell me that you are already committed to the notion that an audio transducer hears -- what else could there be to hearing?

(The same is true for KNOWING rules, as opposed to just being able to perform according to them, by the way, so using that equivocal psychological term begs the question. But even speaking about people rather than desk calculators, it is likewise true that one can KNOW HOW to manipulate symbols according to rules and still not understand what the symbols or the manipulations MEAN, i.e., not understand what one is doing: Searle's argument shows that for language, and I can assure you I can replicate it with a child and arithmetic.)

Stevan

From: Dean Z. Douthat Subject: Motor Grounding

Solving guidance and control problems while avoiding solving equations of motion has been accomplished about 40 years ago by homing guided missile engineers. Proportional navigation guidance and torque balance control, to cite just one of several successful combinations, solve these two problems with simple elegance that nowhere approaches the complexity needed to solve the dual equations of motion for two maneuvering bodies. Such parsimony is essential in throwaway systems. Such parsimony is routine in evolution. The essential processes required for such solutions are available in the sensory, associative and motor equipment of Willie Mays catching baseballs, dogs catching Frisbees, frogs catching horse flies, and horse flies evading tongues. Consciousness is irrelevant. The point is how to solve the problem when the amount of neural machinery is well below that needed even to represent a mathematical model, let alone solve it. The lobster's stomato-gastric ganglion solves the robotic accommodating control problem for six legs with multiple gaits using fewer than 100 neurons.

Previously, I was speaking of analog COMPUTING as opposed to analog DEVICES. The idea of analog computing is to avoid solution of a mathematical model that explains a system of interest by building another system satisfying the same [or usually a simplified] model. This system is then exercised as an analog [or description] of the system of interest. Such approaches are useful when the system of interest is costly, remote or otherwise inconvenient for experimentation or where it doesn't even exist yet as during a system design and for predictive purposes. A few analog computers are electronic; most aren't. A non-electronic example uses soap films as excellent analog computers for certain minimum surface area problems whose mathematical models require calculus of variational approaches for solution.

Cognitive maps, for example, are analog computers using neural processes to model external aspects of the world. These are useful for prediction, rehearsal etc. Instrumental learning [temporal difference learning in engineering terms] extends down to invertebrate levels in the animal kingdom. It can be used by antelopes, for example, as parts of an analog computer to solve such problems as optimal foraging through situated action without recourse to dynamic programming, indeed without knowledge of the underlying Markov processes. Categories are neural analog computers for collections related through "family resemblance" to prototypes. Such analog computing solutions are logically prior to their corresponding [symbolic] explanations and provide bases for cognitive skills that eventually lead to conscious symbolic methods, that is, language, planning, logic etc.

My point simply is that narrow over-reliance on such symbolic methods as logic, linguistics and mathematics for SOLVING PROBLEMS [engineering] as opposed to EXPLAINING NATURE [science] is self-defeating. Good explanations are seldom good solutions. From an evolutionary point of view, symbol crunching approaches to solving problems are weak, late appearing side-effects. In engineering and evolution, efficiency counts; in computer science, a hand held calculator and a Cray are equivalent. If evolution relied on symbol crunching, none of us would be here to discuss it.

Sorry this is so brief but I don't have a lot of time to engage in this sort of thing.

D. Douthat

From: Pat Hayes Subject: Motor Grounding II While reading Nitabach's reply to Douthat, I wondered what both of them would make of the observation, made to me many years ago by Seymour Papert, that the gross behavior of a fielder can be obtained by running the following algorithm to control body movement as a function of the motion of the image of the flyball on the optical array:

if it moves left, move left; if right, move right; if up, move back, if down, move forwards. If its stationary, get ready to catch it.

Obviously this could be improved by talking about how fast its moving, etc., but the essential points are here.

Now, my question to Nitabach is, surely Douthat would be right to say that this isnt solving equations of motion, in no matter how simplified a sense. And, pace Simon's ant, isnt it likely that there are many examples of behavior, especially motor behavior, which arise from relatively simple specifications of how to interact with and respond to a dynamic environment, and depend for their success upon complex features of that environment so that a proper analysis of them will involve complex descriptions of it, but which clearly do not themselves involve processing such a description, even of a highly simplified kind?

But my point to Douthat is, this may not compute trajectories or solve differential equations, but it is, surely, symbolic. It is stated symbolically - indeed , how else could it be stated - and it would be processed symbolically while being run. Perhaps you would want to argue that all these symbols are just our theorists device for talking about the workings of the fielders internal machinery, and have no place inside his head. But if you do, you must come up with an alternative mechanism. Its not enough to talk of analog and sneer at digital, your account has to explain how such algorthims can be valid descriptions of pieces of biological (neural?) machinery which connect the eyes to the muscles without ever computing intermediate 'symbolic' results. And one that can account for how it is that people can change their behavior as a prompt and direct result of gaining new information by any one of many cognitive routes.

OK, just a pebble into the waters.

Pat Hayes

From: Dean Z. Douthat Subject: Tracking Intercept Guidance To: hayes@parc.xerox.com

As you can see by my second message on the subject, you are approaching the same track I am following. Obviously, the equations of motion for the ball cannot be solved as the outfielder has no reasonable access to initial conditions for the ball. Papert has something of the right idea but the actual solution to the tracking intercept guidance [TIG] problem is somewhat more complicated. For example, early in flight the fielder may need to go in even though the ball moves up. This distinction is hardest to perceive if the player is in the trajectory plane of the ball. The most frequently seen error by major league outfielders is to take a step or two in on a ball directly over his head or vice versa. In fact, angular rate of change of the line of sight in inertial space is the key item and proper reaction depends on phase of flight, that is, whether the ball is still rising or has started to fall.

Even this explanation is too symbolic. The missile [frog, hawk, outfielder] doesn’t compute even this simple "algorithm", the symbolic explanation is inserted by intelligent observers for their own understanding and design purposes [cf Harnad’s frequent references to the Hermeneutic hall of mirrors]. As you said, it is necessarily STATED symbolically, there is no other way to STATE anything. But missiles, frogs, hawks, flies, dogs have no need to make statements. Though the outfielder can make statements, he has difficulty stating what he does to catch the ball. The reason for such difficulty is it’s almost all subconscious and sub-symbolic.

In the actual systems, approximations to the key quantities are obtained either by learning [outfielder plus lots of practice] or by being wired in [missile, frog, fly, etc.] using analog computing as I discussed in my second message. For example, none of them have inertial reference frames but all have a means of approximating some important inertial kinematic quantities [using a "slide-back" scheme]. The idea is to avoid symbol crunching by building analogs [electronic for the missile, neural for the others] which behave near enough like the differential equations and related symbolic models for practical purposes. Such problems must have been solved non-symbolically using analog computing means hundreds of millions of years before symbol crunching became possible else symbol crunching would never have happened.

Detailed explanatory accounts of how homing missiles work can be found in publications from the Guidance And Control Information Analysis Center [GACIAC] which is a unit of the Illinois Institute of Technology [IIT] in Chicago. For invertebrates see, for example, Kandel’s "The cellular basis of behavior" and for vertebrates see, for example, Flaherty’s "Animal Learning". These contain extensive references to literature that should satisfy your requirement to explain how nervous systems can solve such problems as TIG. No "intermediate ’symbolic’ results" exist or are needed so, of course, they are not computed. Instead, a neural activation analog is used as descriptive model of the desired process and the model is operated to provide needed predictive guidance and control signals. Such models need not be direct connection between sensor and motor but may also employ "associative" neural clusters. But this certainly doesn’t make them symbolic. Symbolic simulations and analysis of such analog models can explain the propensity for certain types of errors as mentioned above for the outfielder.

You are correct to note that if, say, the second baseman is back-pedaling after a popup and hears the outfielder yelling, "I got it, I got it..." the second baseman will immediately alter his path to avoid the outfielder. This doesn't change the underlying mechanisms for TIG which the outfielder and second baseman share with dogs, frogs, hawks, etc. Evolution, having once solved a problem, almost never throws the solution away. It further refines it and often adds others that can act as overrides, as in this example. This now touches on the symbol grounding problem and my comment that motor grounding may be just as important as transduction. The linguistic symbols "I got it" need MOTOR grounding to have direct and immediate motor effects [and affects] on the second baseman.

D. Douthat

From: miken@ai.mit.edu (Michael N. Nitabach)

It is interesting that both Dean Douthat and Pat Hayes (in their recent postings) have suggested that feedback control could be a general mechanism in biological motor control. Dean brings up the example of guided missile control, a classical application of the then newly elaborated theory of servomechanisms. The unifying principle behind all servomechanisms is the iterated (or continuous) process of modifying an actuator command in the light of some sensory error signal, so as to reduce that error. It is easy to see how Pat's fielding scheme fits into this framework. In this case, the feedback control law is particularly simple: "If it moves left, move left; if it moves right, move right; etc." Now, I agree that this is a possible mechanism for accomplishing the task of catching a fly ball. However, I maintain that this is not the mechanism used by any Major League outfielder worth his six figure salary. Outfielders do not home in on the ball in the sense suggested by Pat; rather, they see the ball in flight, predict (based on some model of ballistic motion) where the ball will land, and move directly to the predicted landing site. An outfielder who was observed to "home in" on the ball in Pat's fashion, would be laughed off the field.

Perhaps this is not convincing enough, so let me try another, related example. That this type of model based, predictive mechanism *must* be available can be seen if we consider a motor behavior closer to the limits of athletic ability. A fly ball is hit into the gap, and the outfielder runs full tilt towards the area where the ball is headed, leaps into the air, and barely makes the catch. How has the feedback loop been closed here? There is no opportunity in this situation for the fielder to "home in" on the ball; there is no sensory feedback during his full tilt run for the ball which tells him if he is on the right trajectory. He must *predict* the point of intersection of his path and the ball's, based on some *model* (possibly learned through experience, possibly innate) of ballistic motion.

Let us turn to Dean's new examples. Anyone who has played frisbee with a dog, knows that a skillful frisbee dog doesn't need to "home in" on the trajectory of the frisbee; it sees the frisbee in flight, predicts its future trajectory, and runs and catches the frisbee--hence the dog's ability to catch a frisbee that it can barely reach running at full speed. I am somewhat surprised at Dean's choice of frogs catching flies as a motor act that could be controlled with a servomechanism. The frog's tongue flick is a paradigm of completely ballistic movements; once the tongue flick is initiated, its trajectory cannot be adjusted by any means. Thus, the frog must *predict* where the fly will be at some time in the future, and *plan* his tongue flick such that his tongue will be at that place at that time. Feedback control is computationally and conceptually simple, and useful in many contexts. However, in cases where the time available for the performance of the movement is not orders of magnitude greater than the time it takes to close the feedback loop, feedback control is useless, and model-based feedforward control schemes must be used.

To summarize and conclude this argument, I have suggested compelling reasons why model based feedforward control *must* be used in at least some contexts. Furthermore, the notions of "plan", "prediction", and "model" have arisen in all of the examples I have considered. It seems clear that these notions require the postulation of (1) internal representations of the dynamics of both the task itself (ball or fly catching) and the mechanical apparatus used to perform the task (human body or frog tongue), and (2) computations over these representations. Dean says, "The point is how to solve the problem when the amount of neural machinery is well below that needed even to represent a mathematical model, let alone solve it." Given the large number of synapses present in even invertebrate nervous systems, I fail to see the relevance of this remark. For example, over half of the cells comprising the nematode c. elegans are nerve cells.

I would like to respond to two of Dean's remarks concerning the analog-digital distinction. First, he says, "This system is then exercised as an analog (or description) of the system of interest." Well, which is it, an analogue or a description? If the concepts "analog" and "description" are allowed to commingle (a miscegenation, in my opinion) in a theory of cognition, then that theory has absolutely no significance vis-a-vis the analog-digital distinction. The heart of this distinction is that analog computers implement just that, *analogues* (*not* descriptions) of the system of interest, while digital devices are limited to symbolic *descriptions* (or representations) of a system.

Second, he says "Such analog computing solutions are logically prior to their corresponding (symbolic) explanations and provide bases for cognitive skills that eventually lead to *conscious* symbolic methods, that is, language, planning, logic, etc." [Emphasis on "conscious" mine] Exactly what is meant by such vague attributions as "logically prior", "provide bases", "eventually lead to"? Ignoring this question, this excerpt suggests to me that Dean is relying too heavily on the conscious-unconscious distinction to provide the intuitive weight behind the theoretical distinctions we are discussing here: analog-digital, symbolic-nonsymbolic, model-servo, logically prior-logically posterior. This just will not do; the conscious-unconscious distinction is completely orthogonal to these others.

- [SH: I agree strongly with the objections in the last two - paragraphs and am likewise interested to hear Dean clarify what he - meant by all this.]

Finally, Dean makes several other statements that I, in the name of biology, feel obliged to respond to. First, he says, "The lobster's stomatogastric ganglion solves the robotic accommodating control problem for six legs with multiple gaits using fewer than 100 neurons." (1) Lobsters have eight legs. (2) The neurons of the stomatogastric ganglion have nothing to do with locomotion; they control the rhythmic movements of the lobster's stomach. Second, he says, "From an evolutionary point of view, symbol crunching approaches to solving problems are *weak*, *late appearing* *sideeffects*." [Emphases mine] This statement is false, through and through. Does Dean consider the use of DNA and RNA as completely symbolic representations of protein sequences a "weak, late-appearing side-effect" of biological evolution? "If evolution relied on symbol crunching, none of us would be here to discuss it." Give me a break; this is just meaningless hyperbole. -- Mike Nitabach

- [SH: Here I must demur, because although the DNA/RNA system seems - clearly digital, and perhaps symbolic, it is also clearly not MERELY - symbolic (and perhaps not even merely digital): It interacts - chemically with growth and other processes (some of them as - analog as a biological process ever gets to be) to actually manufacture - the (likewise nonsymbolic) proteins, etc., of which the organism is - built. In other words, the DNA/RNA code is GROUNDED. If you don't - believe it, try getting a REAL organism out of a symbolic simulation - of the DNA code: The hermeneutic hall of mirrors reigns supreme at - all levels of symbolic discussion... Stevan Harnad.]

From: B Chandrasekaran To: harnad@clarity (Stevan Harnad) Subject: Re: Motor Grounding: Symbols vs. Analogs

A persistent feeling I have had reading the exchanges is that two ideas that need to be separated are often conflated: symbolic vs analog representations and the need for grounding. The DNA/RNA example makes this point clearly. They are symbolic representations, but, as you (Harnad) point out, they are not merely symbolic, but grounded in a very specific way. I believe that there exists a species of interesting symbolic representations with precisely that property -- i.e., they are not symbols in the sense of general purpose Turing machines, but are interpretable only by a special purpose machine whose architectural properties provide the appropriate grounding.

This is not say that the symbolic vs analog representation debate is vacuous. Clearly, it is an empirical issue to be decided on a case by case basis, how exactly a particular information processing system works. But the need for grounding alone does not lead to the rejection of symbolic representation as a possibility. In my view, this conflation is due to the general purpose computer model having run amok in the first decades of AI and cognitive science, and having become in most peoples' minds the only model of symbolic computation.

- [SH: As I suggested in "The Symbol Grounding Problem," hybrid and - "dedicated" symbol systems are no longer really symbol systems at - all, in the strict formal sense, for in a pure symbol system the only - constraint on the symbol manipulations is syntactic: It is based on - the (arbitrary) "shape" of the symbol tokens. In a dedicated "symbol - system" there are added constraints over and above the purely formal - ones. In the case of the RNA/DNA, this has to do with the causal - connection between the biochemistry of the "code" and the - biochemistry of the structures it codes for (as mediated by the - developmental and epigenetic process that generates them); and in the - case of a grounded cognitive system, this has to do with the causal - connection between the symbols and the objects they stand for (as - mediated by the perceptual apparatus that picks them out). - Stevan Harnad.]

Motor Grounding and the Definition of "Symbol System" (Postings from Douthat, Hayes and Harnad)

----

(1) From: Dean Z. Douthat Subject: percept, affect, effect

Thanks for correcting my errors on lobsters, I should stick to eating them. But which is the walking control ganglion then? Anyway, the more legs the merrier; my point is we are down to five or six neurons per leg, pretty impressive efficiency.

Model-based predictive means are also important, especially for "going back on the ball" where LOS cannot be maintained. But tracking is preferred when available and needed in the terminal phase. Note the difficulties when ball tracking is "lost in the sun". I didn't mean to imply frogs were doing the same thing, just a similar thing [exterior ballistics] by similar means. I used the term "description" collectively for such processes as using analog models for navigation, rehearsal, guidance, prediction, control. This was contrasted with "explanation" -- the scientific and usually mathematical counterpart, e.g., geo-kinematic differential equations.

Multiplying motor learning and sensor/effector coordination examples applied to navigation, guidance and control merely reinforces the point that purely symbolic approaches are inadequate; but they offer an alternative for grounding. As stressed by Harnad, the key is causal connections, in this case between internal representations and motor actions.

There is yet a third avenue for grounding -- affective processes. Many of these can be viewed as another form of predictive analog modeling but with internal rather than external results. Thus seeing a threat induces fear predicting possible trauma. Again there is causal grounding between perceptions and anticipated value to the animal.

In summary, I propose a somatic tripod upon which a symbolic platform may rest -- percept, affect, effect.

Dean Douthat

----

(2) From: Pat Hayes Subject: Genetic Grounding

Response to Douthat [not latest posting] and Nitabach

Maybe I should have tried to make my points more clearly. Of course Papert's 'algorithm' for ballcatching would not justify a six-figure salary. It is (and was) intended almost as a joke - it can be written in Smalltalk, but will, if run successfully, catch a ball.

But now imagine making it more realistic, paying attention to angular velocities, phase of the balls trajectory, etc., as suggested by Dean Douthat. Surely this only makes my point more strongly: it still isn't solving differential equations, but it is (even more than before, being now full of references to the current computational state) still a computational specification, the sort of thing which I would call 'symbolic'. Indeed, the more intricate it gets, and the more full of case-analysis (phase of flight, etc), the less plausible is its reduction to the behavior of a completely nonsymbolic mechanism.

One has to be careful about what is meant by 'symbolic'. Any computer program is ultimately compiled into a series of electronic state-changes in a physical mechanism, after all. There is no *sharp* distinction between 'symbol crunching' and any other kind of (physical) mechanism. Douthat talks as though we were discussing turbines versus diesel engines, but this is just a category mistake, in my view. What makes a mechanism essentially symbolic is when its functioning can only be fully explained by treating parts of its internal states as bearers of information which is relevant to the success of that behavior. This is not a different sort of machine, a symbol-cruncher, to be put in a separate category from, say, electronic servomechanisms. On this view, a thermostat is symbolic, albeit in a very trivial way.

I think Douthat has in mind an intuitive distinction which is widely recognised, but not often clearly articulated. And here is where I must take issue with Nitabach, who otherwise I entirely agree with (and who has said much of my reply to Douthat for me). While it is easy to sneer, there is a rather glaring fact which needs to accounted for, which is the extraordinarily rapid evolution of humanity and the apparent correlation between our amazing success - considered simply as a biological species - our brain size, our language, and our conscious awareness. There does seem to be a rather strong a priori case that these facets of humanity are closely coupled together. One view of the link is to insist on a close relationship between ’symbol’ and consciousness and language. A different position, one which computer science leads one towards, breaks this connection and insists that ’symbolic’ is a much more general and abstract notion, so that DNA is already to be understood symbolically. Douthat is I think in the former camp, Nitabach (and I, and many of the correspondents) in the latter. To those on Douthats side, unconscious neural machinery is, more or less by definition, nonsymbolic. Many of the connectionists, and John Searle, among others, would agree.

To me, ours feels like a more coherent position, and one that has more scientific promise. But the other one is not to be dismissed with contempt, as by rejecting it we are faced with the need to give an explanation of awareness, problem-solving, language use, belief, and so forth - all the phenomena usually labelled ’cognitive’ - which goes beyond merely talking of internal symbolisation, of a Language of Thought. If symbols are everywhere, if this way of talking is just a useful approach to thinking about the internal states of animals and molecules, what makes the LOT such a remarkable thing to have in ones head? What is it that evolved so late and was so incredibly successful and seems to need an outreagously large associative cortex to support? People do seem to be *different* from the other mammals, and if one were forced to characterise the difference, to say that we were symbol-users would not seem such a bad second try. (In fact, I think it is essentially correct.)

Pat Hayes

----

(3) From: Stevan Harnad Subject: What are "Symbols" and "Symbol Systems"?

I don’t think it helps to speak of symbols in the extremely general way Pat Hayes does. Here is a semiformal attempt to condense from the literature what does and does not count as "symbol" and "symbol system." It is section 2.1 of "The Symbol Grounding Problem" (Physica D 1990, in press). Without some focus like this, any physical structure or process can be called a "symbol," and then assertions and denials about what can and can’t be done with symbols become so vague and general as to have no consequences one way or the other. What is described below is the kind of system Searle showed to be unable to understand and the kind of system I claim has the symbol grounding problem. If you have something else in mind, we’re not talking about the same thing. (Challenges to these criteria -- which do not amount to an exhaustive list of necessary and sufficient conditions in any case -- are welcome, as are counterexamples and rival definitions.)

1.2 Symbol Systems.

What is a symbol system? From Newell (1980) Pylyshyn (1984), Fodor (1987) and the classical work of Von Neumann, Turing, Goedel, Church, etc.(see Kleene 1969) on the foundations of computation, we can reconstruct the following definition:

A symbol system is:

(1) a set of arbitrary PHYSICAL TOKENS (scratches on paper, holes on a tape, events in a digital computer, etc.) that are

(2) manipulated on the basis of EXPLICIT RULES that are

(3) likewise physical tokens and STRINGS of tokens. The rule-governed symbol-token manipulation is based

(4) purely on the SHAPE of the symbol tokens (not their "meaning"), i.e., it is purely SYNTACTIC, and consists of

(5) RULEFULLY COMBINING and recombining symbol tokens. There are

(6) primitive ATOMIC symbol tokens and

(7) COMPOSITE symbol-token strings. The entire system and all its parts -- the atomic tokens, the composite tokens, the syntactic manipulations (both actual and possible) and the rules -- are all

(8) SEMANTICALLY INTERPRETABLE: The syntax can be SYSTEMATICALLY assigned a meaning (e.g., as standing for objects, as describing states of affairs).

According to proponents of the symbolic model of mind such as Fodor (1980) and Pylyshyn (1980, 1984), symbol-strings of this sort capture what mental phenomena such as thoughts and beliefs are. Symbolists emphasize that the symbolic level (for them, the mental level) is a natural functional level of its own, with ruleful regularities that are independent of their specific physical realizations. For symbolists, this implementation-independence is the critical difference between cognitive phenomena and ordinary physical phenomena and their respective explanations. This concept of an autonomous symbolic level also conforms to general foundational principles in the theory of computation and applies to all the work being done in symbolic AI, the branch of science that has so far been the most successful in generating (hence explaining) intelligent behavior.

All eight of the properties listed above seem to be critical to this definition of symbolic. Many phenomena have some of the properties, but that does not entail that they are symbolic in this explicit, technical sense. It is not enough, for example, for a phenomenon to be INTERPRETABLE as rule-governed, for just about anything can be interpreted as rule-governed. A thermostat may be interpreted as following the rule: Turn on the furnace if the temperature goes below 70 degrees and turn it off if it goes above 70 degrees, yet nowhere in the thermostat is that rule explicitly represented.

Wittgenstein (1953) emphasized the difference between EXPLICIT and IMPLICIT rules: It is not the same thing to "follow" a rule (explicitly) and merely to behave "in accordance with" a rule (implicitly). The critical difference is in the compositeness (7) and systematicity (8) criteria. The explicitly represented symbolic rule is part of a formal system, it is decomposable (unless primitive), its application and manipulation is purely formal (syntactic, shape-dependent), and the entire system must be semantically interpretable, not just the chunk in question. An isolated ("modular") chunk cannot be symbolic; being symbolic is a combinatory, systematic property.

So the mere fact that a behavior is "interpretable" as ruleful does not mean that it is really governed by a symbolic rule. Semantic interpretability must be coupled with explicit representation (2), syntactic manipulability (4), and systematicity (8) in order to be symbolic. None of these criteria is arbitrary, and, as far as I can tell, if you weaken them, you lose the grip on what looks like a natural category and you sever the links with the formal theory of computation, leaving a sense of "symbolic" that is merely unexplicated metaphor (and probably differs from speaker to speaker). Hence it is only this formal sense of "symbolic" and "symbol system" that will be considered in this discussion of the emergence of symbol systems...

Stevan Harnad

[To Recipients of the Symbol Grounding Discussion List: Please let me know if you want your name to be removed from this list. Stevan Harnad harnad@clarity.princeton.edu harnad@pucc.bitnet]

From: Eric Dietrich

The Chinese Room and Abortion

I originally wanted to present an argument that the systems reply is correct. I now despair of ever being able to do this. The reasons are given below. Before I get to them, however, I would like to correct a misunderstanding. One argument of Harnad's against the systems reply is that the system is Searle, period. Since Searle is doing everything in the system, and Searle is not understanding, the system isn't. Therefore, the system reply fails.

This particular argument of Harnad's misses the point of the systems reply. To the true systems-replier, the fact that Searle doesn't understand Chinese is irrelevant. Yes, Searle-engaging-in-Chinese- Room-behavior is all there is to the system, and yes Searle does not understand Chinese. But the systems-replier is prepared to see many systems co-occurring in one place. The relevant system is Searle-engaging-in-Chinese-Room-behavior, and this system does understand Chinese... This system has intentionality... This system probably has rights. Etc. Etc. In short, Searle is not the system of interest. Searle is a system, but not the one under discussion. Harnad's argument turns on conflating Searle with the Searle-Chinese-Understander system.

This much is obvious to me. Several weeks ago, I thought that were I to point out this tiny error, Harnad and others would change their minds. Now, I don't believe this. (I could be wrong. Perhaps now Harnad and Searle and others are saying, "Of course. Thanks Eric. How about tenure at Harvard." I'll assume they aren't saying this.)

Something is going on beneath the surface. Searle’s argument is justly famous. Perhaps it is one of the 20th century’s great puzzles. But the disproportion between the amount of discussion, and the amount of movement on either side is the greater puzzle. Most of the discussion regarding the Chinese Room amounts to repeated assertions that the systems reply is correct, and repeated denials that it isn’t. Those on both sides wonder how their otherwise intelligent colleagues can be so misguided. What is obvious to one side, is absurdity to the other. The question therefore is "What is really at stake, here?"

Consider the abortion issue. A pro-lifer and pro-choicer whose debate goes no deeper than "Can’t you see that abortion is wrong?" and "Can’t you see that abortion is a woman’s right?" will get nowhere. The crucial issue for them is what is beneath the surface. Beneath the surface, pro-lifers and pro-choicers have completely different agendas. The agendas are vast, and derive from different central priorities. Roughly, for the pro-lifer, the central priority is the sanctity of human life (for whatever reason). For the pro-choicer, the central priority is the right of individuals to control their destinies (as much as possible). These priorities are not logical opposites, but they are set against each other in the abortion arena. Something like this is going on in the Chinese Room debate.

For many years, I thought that the anti-systems types were really closet dualists, and I have said so in print. This would explain much. The anti-system types are dualists: they think humans are not mechanisms. Systems-repliers are mechanists (physicalists): they think that humans are machines. But the anti-systems types repeatedly deny being dualists. Perhaps they are speaking the truth. But if they are not dualists, then something quite odd is going on. How can mechanists disagree so completely?

I suspect that the anti-system repliers have a completely different agenda than the systems repliers. Arguments that presuppose these agendas are going to be unpersuasive. And all of the arguments I’ve seen for the last 10 years presuppose these agendas. To make any progress, therefore, we must debate the issues beneath the surface. Unfortunately, I’m not sure what the different agendas are. Even worse, if the different agendas presuppose different metaphysics (and I suspect they do), then all discussion is moot. It is a well-kept secret of philosophy that metaphysical debates are worthless.

In my own case, I know my deeper agenda. I seek an explanation of human behavior that makes it continuous with cockroach behavior, tumbleweed behavior, and rocks-heating-up-in-the-sun behavior Anti-systems repliers seem to me to want an explanation of human behavior that exhibits how humans differ from cockroaches, tumbleweeds, and rocks heating up in the sun. I agree that humans are different than cockroaches, tumbleweeds, and warming rocks. But their similarities are more important. Moreover, if we ever want to explain the differences between humans and cockroaches, we must first understand in detail their similarities. The differences between humans and cockroaches are going to turn out to be variations on a theme. We must first find out what this theme is. This is the intuition behind the systems reply.

Since I’m starting to wax metaphysical, I’ll stop.

Eric Dietrich

TO THE "TRUE SYSTEMS-REPLIER": ON THE RIGHTS AND WRONGS OF SYMBOL SYSTEMS

From: Stevan Harnad

I discern no argument whatsoever above, just another reflexive repetition of the original hypothesis ("if a system passes the TT, it MUST understand") in the face of the evidence that refutes it (the only system in sight, Searle, does NOT understand), with some suggestions that to think otherwise is to commit metaphysics... It seems to me that the only thing this demonstrates is the vivid and tenacious sci-fi fantasy life of the seasoned AI Trekkie. After all, one can hardly expect to change someone's mind once it's made up (like the little boy, the pile, and the proverbial pony) that there's got to be another mind here someplace...

Yet perhaps there's still something to be said for preferring to cling to the remnants of one's common sense, rather than freely multiplying entities just because of a dogged preconception that has been borne aloft and out of reach by a flight of fancy:

One would (wouldn't one?) rightly have a bit of a problem before a court of law, for example, in trying to argue for the rights of "the system" -- i.e., Searle-engaging-in-Chinese-Room-behavior -- as distinct from Searle, if the only entity before the bench was Searle (who already has his rights).

Calling the law's unwillingness to believe in dual personality under THESE conditions "dualism" is a bit of a desperation measure. The law tends to be quite monistic, materialistic even. It would have no problem, in principle, with the rights of a cockroach, perhaps even a rock; I can even conceive of distinct sentences for the different personalities allegedly residing in a person with multiple personality disorder. (In other words, under the right conditions, the law too, like "the systems-replier, is prepared to see many systems co-occurring in one place.")

But before you rush to enter that plea in the case of Searle vs. Searle-engaging-in-Chinese-Room-behavior, I must warn you that expert witnesses can be called -- specialists on multiple personality disorder -- who are prepared to testify that memorizing a bunch of symbols is NOT among the known etiologies of the disorder, nor is it compatible with any of the known facts about brain function or psychological function that it will ever turn out to be.

In any case, the charge of dualism would never hold up in court, because I've already proposed a perfectly nondualistic alternative for the refuted hypothesis that thinking is only symbol crunching: Thinking must consist of nonsymbolic functions too -- which, may it please the court, is still perfectly mechanistic, and could occur in cockroaches, perhaps even in tumbleweeds...

Clarence Darrow, Q.C.

P.S. I've made these and many other points many times before. Perhaps the "disproportion between the amount of discussion, and the amount of movement on either side" noted by Dietrich may result from the fact that what is going on is not just "repeated assertions that the systems reply is correct, and repeated denials that it isn't." There are also arguments and evidence that it isn't, which are repeatedly ignored, as the systems reply keeps getting parroted back, at a greater and greater counterfactual price, by the "true systems-replier," whose relenteless sci-fi commitments seem to amount to a "different agenda" indeed. (The differences, it seems to me, have less to do with the metaphysical than the supernatural.)

We can, after all, agree that "the differences between humans and cockroaches are going to turn out to be variations on a theme," and that "we must first find out what this theme is." But we must also be prepared to accept the verdict that this theme has turned out NOT to be symbol manipulation. That's just open-minded empiricism...

Stevan Harnad

From: Dean Z. Douthat Subject: neo-Pythagoreans

Steve Harnad has basically stated my position as to what counts as a symbol system and the difference [per Wittgenstein] between explicit and implicit rules. You are correct in assuming I don't regard a neuron [or most collections of them] as a symbol system but not by definition at all. Rather because it fails to meet several of the criteria set out by Harnad.

The failure to make the distinction between explicit and implicit rules [symbols] has serious epistemological drawbacks as you have already noted. Just as serious are the metaphysical difficulties. Where is the floor of symbol systems? Is the soap film on a wire frame a symbol system? After all, it "solves" intgro-differential variational equations. This neo-Pythagorean position would have not only Willie Mays as a symbol system but also the ball [and each atom therein [and each hadron therein [and each quark pair, lepton, neutrino qradruple therein [and ... Few physicists [e.g., Eddington, Kaplan] are willing to follow this path to the end and they are not seriously regarded qua neo-Pythagoreans.

As to the explosion of associative cortex, I do believe this is correlated with emergence of explicit symbol manipulation capability, specifically language and logic. Along with Dennett [Daniel Dennett "Brainstorms: Philosophical Essays on Mind and Psychology", MIT Press, 1981], I believe neurons support this capability by building a "virtual serial processor" to handle explicit symbol crunching. So I regard the particular collection(s) of neurons supporting these activities as instantiating a symbol system. Since neurons are inefficient at this kind of processing, it takes a lot of them, thus the cortical explosion.

The "amazing success" of the human experiment remains to be seen; check back in 10 million years.

D. Douthat

Dan Hardt hardt@linc.cis.upenn.edu University of Pennsylvania wrote:

> I'm not sure how you can sharply distinguish between a system
> that is interpretable as rule-governed and one that is
> explicitly rule governed. Perhaps you have in mind a connectionist
> network on the one hand, where what is syntactically represented might
> be things like weights of connections, and the rules only emerge from the
> overall behavior of the system; on the other hand, an expert system,
> where the rules are all explicitly written in some logical notation.
> Would you characterize the connectionist network as only interpretable

> as being rule-governed, and the expert system as being explicitly
> rule governed? If it is that sort of distinction you have in mind,
> I’m not sure how the criteria given allow you to make it. If fact, I
> wonder how you can rule out any turing machine.

I’m willing to let the chips fall where they may. All I’m trying to do is settle on criteria for what does and does not count as symbol, symbol system, symbol manipulation.

Here is an easy example. I think it contains all the essentials: We have two Rube Goldberg devices, both beginning with a string you pull, and both ending with a hammer that smashes a piece of china. Whenever you pull the string, the china gets smashed by the hammer in both systems. The question is: Given that they can both be described as conforming to the rule "If the string is pulled, smash the china," is this rule explicitly represented in both systems?

Let’s look at them more closely: One turns out to be pure causal throughput: The string is attached to the hammer, which is poised like a lever. Pull the string and the hammer goes down. Bang!

In the other system the string actuates a transducer which sends a data bit to a computer program capable of controlling a variety of devices all over the country. Some of its input can come from strings at other locations, some from airline reservations, some from missile control systems. Someone has written a lot of flexible code. Among the primitives of the system are symbol tokens such as STRING, ROPE, CABLE, PULL, HAMMER, TICKET, BOMB, LOWER, LAUNCH, etc. In particular, one symbol string is "IF PULL STRING(I), LOWER HAMMER(J)," and this sends out a data bit that triggers an effector that brings the hammer down. Bang! The system also represents "If PULL STRING(J), LOWER HAMMER(J)," "IF PULL STRING(J), RELEASE MISSILE(K)," etc. etc. The elements can be recombined as you would expect, based on a gloss of their meanings, and the overall interpretation of what they stand for is systematically sustained. (Not all possible symbol combinations are enabled, necessarily, but they all make systematic sense.) The explicitness of rules and representations is based on this combinatory semantics.

It is in the latter kind of symbol economy that the rule is said to be explicitly represented. The criteria I listed do allow me to make this distinction. And I’m certainly not interested in ruling out a Turing Machine -- the symbol system par excellence. The extent to which connectionist networks can and do represent rules explicitly is still unsettled.

Stevan Harnad

[Recipients of the Symbol Grounding Discussion: Please let me know if you wish your name removed from this list. -- Stevan Harnad]

The following message from Michael Dyer is one with which I basically agree. Careful readers will note that I have pointed out all along that it could well turn out that connectionism is just a form of symbolism (although I don’t think the question is altogether settled yet). The symbolic/nonsymbolic matter would then be a separate one, and a deeper one.

I will accordingly not respond to this posting from Mike; but below it I excerpt from and respond to another one (which I have not reposted in its entirety) on the symbolic/nonsymbolic matter.

Stevan Harnad

From: Dr Michael G Dyer Subject: symbolism vs computationism vs intentionality

To everyone out there on Harnad’s mailing list, a general comment:

It is my understanding that both Searle’s and Harnad’s arguments against intentionality in the chinese box have nothing at all to do with the debate currently going on between symbolists and connectionists. The symbolist/connectionist debate is not about which approach is fundamentally more powerful (since they’re equivalent) but which approach is a more natural way of characterizing cognition (or different aspects of cognition).

We should not let the symbolist/connectionist debate split all the "systems" people into thinking they are on different sides of the chinese-box debate.

Searle’s thought experiment does not need to specify exactly WHAT those chinese box instruction books specify: they could be telling the person in the box to place a chinese character on top of a giant matrix of numbers and then have the person do matrix operations to simulate movement of signals from a simulated ’retina’ to other topographic "maps", to simulate neuron pulse firing, thresholds, etc. Searle would STILL claim that there is no intentionality because the PERSON in the box has memorized the instructions and still doesn’t experience understanding chinese. So the issue of whether or not it’s prolog, or frames, or artificial neural nets or spreading activation is irrelevant to Searle’s argument.

Harnad’s argument is ALSO orthogonal to the symbolism vs connectionist debate. Like Searle, his attack is on the general notion of computationism (of which pure symbol processing and PDP-style processing are both examples). Harnad argues that intentionality needs what he calls "nonsymbolic" processing, such as transducers for sensory/motor behaviors in the real world. Harnad (rightly) equates "symbolic" processing with "computational" (since a turing machine is "symbolic" and can simulate any computation). So Harnad is attacking the notion of computation in general, arguing that one cannot have intentionality unless one has noncomputational operations in the system. For example, the movement (in the real world) of a person, from loc x to loc y, is normally not considered a computation. Harnad argues that without all this noncomputational "stuff" one cannot have intentionality and that this stuff is more fundamental than computation and that the totally simulated robot (i.e. all inside a machine) is not intentional because there aren’t these noncomputational things going on.

Now, if people want to argue over various forms of symbolism vs various forms of connectionism, and their relative problems and merits, that’s fine. But it really has nothing to do with the Searle/Harnad arguments and all of you people out there who believe that mind/intelligence/perception/ "intentionality" are the result of the COMPUTATIONS executed by some physical system, should speak up and address the Searle/Harnad position (because I’M getting tired of "holding the fort").

Michael Dyer

From: Stevan Harnad

On an earlier, Mike Dyer wrote:

> I have shown these debates to several of my colleagues and they really
> have a mocking attitude toward the Harnad/Searle arguments. You really
> have a mocking attitude toward the "systems" crowd, that it's wild sci
> fi, absurd etc.

I agree that what one mocks is a matter of taste (sometimes bad taste). But the reasons I find your position unconvincing are quite explicit. Do you really think that I'm the only one who would describe what follows below as sci-fi?

> suppose every now and then Searle suddenly started running around,
> speaking chinese (we'll generalize from reading) and manifesting a
> completely different personality. suppose the chinese-searle was making
> investments etc, with "english-searle"'s money, etc.
>
> suppose that english-searle was manifesting itself less and less, until
> english-searle hardly every "appeared". Now we all end up in court
> and chinese-searle gives eloquent debates (with the aid of a chinese
> interpreter)... better yet, assume that chinese-searle takes over the
> one body, while traveling in taiwan, so we end up in a chinese court
> anmd english-searle needs an english-to-chinese interpreter...
>
> chinese-searle will argue that he is the dominant one, since he is
> manifest 95% of the time. sure, english-searle will give the arugment
> that he is doing calculations of some memorized book to bring about
> chinese-searle, but what court is going to believe THAT story?
>
> My point is, that we should try NOT to rely on sarcasm and other rhetorical
> devices and instead, give our arugments in as straightforward a manner.

I agree that we should be polite and respectful of one another, but sometimes you just have to call a spade a spade. Is there a value-neutral term that could describe the above? Extended conjecture?

> Stevan, earlier I stated your argument in my own words and you stated
> that I had understood it. I then gave a rebuttal and you seem
> unconvinced by it. Just as you can ask why it is that the systems
> people are unconvinved by your and searle's arguments, the systems
> people can ask why you seem unconvinced by the "split personalities"
> rebuttal... Your rejection of this rebuttal seems to rely on what is
> currently known about humans with split personalities -- that is
> unfortunate since we are really not talking about these controversial

> people, but about /##/ the extent to which multiple virtual systems can
> co-inhabit the same hardware, which is rather common in modern
> operating systems. I think your rejection of this rebuttal gets back to
> how you (and searle) imagine feeling as you try to be one of the
> virtual systems... but we know we can have a lisp system create an AI
> system and it creates it by doing all aspects of what is needed to bring
> the AI system about; yet, if we talk to the lisp system, it will not be
> able to tell us anything about the AI system to find out what it's like
> to be the ai system, you'd have to talk to it.
>
> IF the disagreement hinges on notions of state of consciousness,
> qualia, and imagining the consciousness of others, then it IS
> hopeless...

Mike, I agreed that you had paraphrased my argument correctly; but paraphrase is not the only measure of whether an argument has been grasped. What you took to be a rebuttal is partly an indication of what you took to be the argument, and the reason I am not convinced by your rebuttal is that misses the point: Please take it just as a piece of imagery and not a sign of disdain that I say that it is at the point I have marked by "/##/" that I find you seem again to have taken off into the twilight zone. Discussing the real conditions of real multiple personality disorder is one thing. Freely conjecturing that a radically different condition, for which there is no human evidence whatsoever, could also cause it in human beings because of certain human-like interpretations you have projected on your computer just is not a counterargument. At best, it is trying to answer negative evidence by conjecture, and that doesn't work, logically. That's why I keep saying you're taking out a bigger and bigger counterfactual debt.

And the issue has nothing to do with controversial notions about consciousness. Right now, in 1989, based on everything we know, whether in a court of law or in a refereed scientific journal, the following can be stated with as much certainty as any other empirical fact: There is no evidence whatsoever for believing that memorizing rules, no matter how many or how complicated, can or ever will cause multiple personality in any human being. Period. That's the only commonsense fact that Searle's thought experiment draws upon. No weird ideas about qualia...

Stevan Harnad

Mike, this is not being posted to the group:

> Come on Stevan! What makes you think that Searle doing all that
> imagining of carrying out all those (unspecified) instructions and then
> stating he does or does not sense the "understanding" is any more "out
> of the ordinary" than MY conjectures about a Chinese court listening to
> the Chinese-searle argue for it's own "intentionality" being more
> primary than the English-Searle's???? What makes YOUR or SEARLE's
> imaginings OK, but MINE the butt of your incredulity???

I give up. Simple induction on real life symbol memorization allows you to extrapolate to where it is likely to lead if you did a lot more of the same. No fantasy needed -- but the notion of Searle gradually being taken over by his alter ego under those same conditions...? If you don't see the difference, what can I possibly say?

> Also, I'm tired of you having the LAST word on each of MY messages.
> Each time YOU send a message, it appears in its entirety and people
> get to read it as such, and THEN they reply. But each time *I* (or anyone
> else) sends a message, it is NOT sent out UNTIL after you have
> peppered it with all of your comments of incredulity. I don't find
> it an argument on YOUR part that you find something credulous. That gets
> back to my point about seeing if we can figure out the deeper beliefs
> of each group.

I don't annotate every message. But, as I said several iterations ago, our exchange has already gone full circle several times. If there is one last message you want me to transmit without annotations, I'll do it, but I really think that both sides have had a chance to say everything they're able to say, and nothing new is flowing any more. Perhaps eventually someone else will say something that will again stimulate one or the other of us to come up with something genuinely new on our points of disagreement. But for now it's gone stale.

> Go back and look at YOUR comments on my arguments -- they're mostly
> rejecting my arguments, based on the assumption that your and Harnad's
> arguments are definitive, or a rejection based on simple incredulity,
> and no principled response to my arguments...

I was under the impression that I did give a reasoned response: It was that your responses to Searle are not arguments, they are conjectures.

> What makes you so sure that SEarle (and YOU, and ME) are not already
> amoeboid forms of consciousness (a la Minsky's "Society of Mind")?

I'm not sure; I don't have to be sure that every far-fetched conjecture is false. I just have to be satisfied that its probability -- like the probability that there's a God, special creation, or leprechauns -- is too small to take seriously.

> When two groups cannot convince each other, then there are only two choices:
> 1. quit talking (which is not an unreasonable strategy)
> or
> 2. seek the deeper assumptions

Both of those are fine; what's not fine is just repeating what one has said before.

> [1] why is it ok to have virtual systems operate on the same computer
> hardware but deny that possibility in brains? [2] What do you think
> Searle's brain must be doing in order to memorize and automatically
> carry out the chinese-box instructions? [3] If the chinese-searle can pass
> turing's test, then it sure CAN argue for its own rights in the chinese
> court of law.

[1] Because I don't think a "virtual system" in a computer and a mind in a brain are the same sort of thing. [2] I think Searle's brain must be doing the same thing I do when I memorize and carry out meaningless symbol manipulating instructions: meaningless symbol manipulation. [3] I think a court of law would rightly deny that there were two people before the bench when Searle did his Chinese act; and if it didn't, I still think it would be wrong, and Searle would be right.

Now Mike, I've answered you again, this time just between you and me, but I'm afraid I can't continue to do this. There is definitely an obsessive side to the Chinese Room Problem, and I'm as susceptible to it as anyone else. But I really have heard your full repertoire on this; if you want to continue airing these same points you'll have to start a discussion group of your own (and then YOU may find that you'll have to declare a cease-fire with one of your interlocutors because the exchanges have gotten too repetitious!).

I don't want to be dictatorial, but part of the purpose of these discussions was to demonstrate the potential of scholarly skywriting in the development and communication of ideas. So I really have to exercise the judgments and prerogatives of someone who has been an editor for a decade and a half to moderate the flow of the discussion. Our cease-fire was declared several iterations ago: Can we observe it now?

Peace,

Stevan

To: geller@vienna.njit.edu (james geller)

Jim, I'd be happy to talk at NJIT. Send me several possible dates, as I'm giving a lot of talks in the next few months and it will take a little juggling to pick a window that's open.

> Is there a mailing list at Princeton, where we could publicize
> Rosenfeld's talk?

Send it to me and I'll post it for you.

> As I told you, I gave a presentation at a workshop, where I briefly
> mentioned your work (in answer to a question). Some of the feedback
> that I received made me realize that what I am advocating is really a
> form of "propositional grounding". One term is grounded by using it in
> a definitional sentence with other terms that are already grounded.

That's correct. But remember that definitions are circular unless the definienda are grounded, and that grounding must eventually go down by recursion to an elementary grounding that is direct sensory rather than definitional. It's the initial sensory grounding that's critical. (Do you have references for "propositional grounding"? Who uses it, and how?)

> The group at Buffalo is very implementation oriented, and some feedback
> made me remember that I should start to implement some of these ideas in
> a parser. (Please do not spread this, I want to get at least a start on
> it, before I tell other people).

Parsers come AFTER sensory grounding, not before, or instead...

> The other thing I realized is that perceptual grounding, the way you
> suggest it, is probably quite hard for "form words." It is also much
> easier for spatial terms than for temporal and abstract terms. I
> suspect that there is a "gray area" between perceptual/motoric
> /emotional grounding and propositional grounding. I need to think about
> this more, again, please keep it confidential for a while.

Grounding is a bottom-up procedure, starting at the concrete, sensory level. Abstract words are grounded in concrete ones, and the most concrete ones must be grounded directly in sensory category detectors. (Don’t worry so much about have ideas stolen. Separate little ideas never amount to anything. And a creative thinker should have plenty more where that came from. Let your ideas flow freely, seek feedback, and build up a bigger systematic set of ideas that is so uniquely your own that you have bigger worries about getting people to understand their structure as clearly as you do than about anyone walking off with it!)

> One last thing, I am several messages behind reading your mails, but did
> you ever talk about experimental verification of the symbol grounding
> hypothesis? It seems to me that propositional grounding might be amenable
> to experimental verification. Another thing I just started to think about,
> and that needs work.

I have a collaborative project with some experimentalists at Vassar in which we are testing whether we can get a learned categorical-perception boundary effect with simple sensory continua. These would be the most elementary categories. I am also collaborating with some neural net simulations of learned categorical perception effects with Steve Hanson at Siemens.

I advocate bottom-up grounding, but in principle, a recursive procedure for generating new grounded symbols out of already grounded ones could be investigated directly. Recall, though, that I am rejecting symbolic modularity, because I think there will be an important constraint on the higher order symbols that is inherited from the fact that the ones at the ground level are causally connected to nonsymbolic processes that pick out the objects they refer to from their sensory projections. This constraint is missing in a purely top-down module that inherits grounding by mere recursion.

Stevan Harnad From: Pat Hayes

Come come, Stevan, stop spluttering. Let me try to state some of the systems reply in a way which might provoke a calmer reaction. I am sure you still wont agree, so Im not sure why Im bothering, but I will try.

My first point about Searles rather desperate (as it seems to me) response to the systems reply can be illustrated very nicely by going into your hypothetical court of law. This fellow speaks perfectly good Chinese half the time, but the rest of the time says he cant understand a word of it, exhibits selective amnesia, and says he is running a computer program in his head. When you talk to him in Chinese he has no memory of the times when he is talking English, in fact is not even aware of that aspect of his personality. He is clearly crazy. Anyone who exhibited the behavior which the Searle system is supposed to would be immediately declared unfit to plead by any reasonable court, and put under psychiatric care.

The point of this is that this kind of example is just misleading as an ’intuition pump’. For example, you take it as just obvious that there is only one system there - ’the only entity before the bench was Searle’ - but actually, just in order to keep track of what was going on in the court records, all the people iteracting with him would indeed be obliged to introduce the entity ’Searle-speaking-Chinese’ and distinguish it from Searle-speaking-English, and this distinction would be both intuitively compelling and essential for legal and scientific communication. So why would it be so obvious that there was just one ’system’ present? In fact, I think that one would either come to think that there were two personalities inhabiting that body, or else one would conclude that Searle was lying when he said he didnt understand Chinese; and that this would be what most of the legal argumentation would be about.

(Its true that it would be hard to argue for the rights of one of the personalities, but that is because the law does not recognise such fragmented personalities as distinct individuals. Thats a legal matter, however, not a philosophical one. The law is currently notoriously not able to handle questions of individuation to which software, even of a mundane sort, gives rise. And Im not very persuaded by your appeal to the authority of ’specialists in multiple personality disorder’, either.)

But all of this kind of discussion is silly, because Searle couldnt memorise and run a computer program like the one that would be needed to pass a turing test. And thats a matter of principle, not mere pragmatism. Just the speed is impossible, for a start. But apart from questions of complexity, suppose the judge asks Searle - not the room, but Searle while he is running the room program - (the Chinese for) "Whats in your inside jacket pocket?" Is Searle going to open his coat and look inside, all the while chatting in Mandarin? If so, what is the American Searle making of this: is this program which he is running in his head taking control of his body? Suppose one tells him, in Chinese, to roll his eyes, or scratch his nose? Or asks him whether he is hungry or has a headache? The list can be extended indefinitely. By putting the room program into a body, one has made it necessary to attach the code to that body, in order to keep the example sensible. Similar points can be made in connection with memory: if the Chinese fellow remembers things from one day to the next, does the American also come to know them? How? And if not, then surely this points even more strongly to the separation of the two systems in this poor overburdened head.

I imagine that you, Stevan, might find these objections quite congenial. After all, what Im saying is that in order to for the example to work, some of the symbols in the Chinese Searle must be grounded in his physiology. But of course: I suspect that most proponents of the symbolic approach to AI and the systems reply to Searle would not disagree for a moment. But notice that the example needs to ASSUME this in order to be a convincing argument FOR it, although this assumption is usually not brought out into the open, and that it is the step from the Room to Searle-having-memorised-the-code, the standard reply to the systems reply, which makes it necessary. (After all, one would not expect a room to have inside pockets, would one?)

You accuse the systems repliers of taking out a greater and greater counterfactual loan, but let me turn that accusation back at you (and Searle): the whole silly discussion, taken as an objection to the symbolic hypothesis, relies on counterfactuals which are so outrageous that it is remarkable that anyone has taken them seriously for more than five minutes. The whole argument is simply an appeal to a naive intuition based on a failure to grasp some of the basic ideas of computer science, and relying on claims about what would be ’obvious’ in circumstances which are impossible, and which if one tries to take seriously for more than a few moments suggest quite different intuitive conclusions. I enjoy these arguments as much as the next man (and to be fair, this discussion has led to us all sharpening the precision of our vocabulary) but to draw serious methodological or scientific conclusions from this stuff is - putting it politely - to misplace ones confidence.

Pat Hayes

Reply to Hayes: From: Stevan Harnad

> hypothetical court of law... selective amnesia... clearly crazy...
> ’Searle-speaking-Chinese’... Searle-speaking-English...
> one would either come to think that there were two personalities
> inhabiting that body, or else one would conclude that Searle was lying
> when he said he didnt understand Chinese... I’m not very persuaded
> by your appeal to the authority of ’specialists in multiple personality
> disorder’

Ah, but you haven’t described the real situation in this hypothetical court of law at all! What we really have is Searle, compos mentis at all times, no amnesia, truthfully telling the court that he does not understand Chinese (and calling witnesses to corroborate that fact), and that all he had done was to memorize a bunch of symbols and symbol manipulation rules that he likewise does not understand, and that when someone flashes a Chinese cue card to him (or even -- for dramatic value, and since the real Searle actually happens to be an excellent phonetic mimic -- whenever someone speaks to him in Mandarin), all he does (at will, recall, not in some sort of somnambulistic trance!) is respond according to the meaningless rules he’s learned, without the faintest idea of what he’s saying.

Now in Chinese he may be saying "I understand" (and who knows what he’s supposed to be saying about his alter ego -- that depends on his Chinese script), but it seems obvious that there’s nothing in this game of "Simon Says" that would suggest SEARLE was crazy: He has a perfectly good alibi; and moreover, IT’S COMPLETELY TRUE! And the specialists are right, whether or not you are more persuaded by them than the sci-fi overtones you are needlessly projecting onto an otherwise perfectly commonsense scenario (apart from its counterfactual premises, to which we will return).

After all, there IS a third alternative, isn’t there? -- Or have we gotten so lost in the hermeneutic hall of mirrors as to have forgotten it altogether? Besides the possibility that Searle really does have two minds, or that he’s lying about not understanding Chinese, there’s still the possibility that he really has only one mind, that he really doesn’t understand Chinese, and that the only thing he’s doing is manipulating meaningless symbols...

This issue has NOTHING WHATSOEVER to do with any "failure to grasp some of the basic ideas of computer science," which are, and ought to be, completely neutral about the mind! What Searle has resisted, and the systems-repliers (who are really just desperate Strong-AI-savers, in the face of negative evidence), have utterly succumbed to, is the apparently irresistible temptation to interpret both code and computer performance mentalistically. But please don't cite "computer science" as the authority on which they do so! Computer science says (and can say) NOTHING about what it is like to be a system executing a code! It never ceases to amaze me how this simple and correct fact keeps being forgotten by believers in Strong AI (aka Systems-Repliers).

> But all of this kind of discussion is silly, because Searle couldnt
> memorise and run a computer program like the one that would be needed
> to pass a turing test. And thats a matter of principle, not mere
> pragmatism. Just the speed is impossible, for a start... complexity...

Yes indeed, it may be counterfactual to suppose that Searle could memorize and execute all those complicated symbol manipulations fast enough. But that's not where the counterfactuality began: It began with the premise -- which Searle accepted, arguendo, from "Strong AI" -- that the Turing Test (the purely linguistic one: only language in and language out), the TT, could be passed by symbol manipulation alone. THAT may well be counterfactual, but we have agreed to swallow it for now, to see where it will get us. But then when Searle wants to imagine whether manipulating symbols really does amount to understanding, we want to stop him because it would be counterfactual to suppose he could do it fast enough, or remember it all!

The logic of this seems a little like supposing that the shroud of Turin was really the robe of the lord, but forbidding us to cut it to test it, because if it were the robe of the lord it would be sinful to cut it. Self-consistent in its own way, but leaving one a trifle dissatisfied...

But look, these quantitative arguments are hand-waving in any case. I don't have any reason to believe in phase transitions in processing speed or capacity: Let's memorise a FEW Chinese symbols and see if we understand Chinese a LITTLE, or exhibit MILD multiple personality disorder. If not, why should I (counterfactually) believe that there's something dramatic waiting to happen down the road past where I can ever reach, when all indications are that it's just more of the same? (This is also typically the way AI handles the problem of "scaling" up from toy performance to lifesize performance, by the way.)

Bottom line: AI is in no position to complain about counterfactuals here; and, unless there is evidence that I haven't heard yet for a phase transition toward the mental at a critical symbol manipulation speed and capacity, extrapolating from perfectly feasible human symbol manipulation performance and experience seems to be a sufficiently reliable inductive guide for present purposes.

> "Whats in your inside jacket pocket?" roll [your] eyes, or scratch
> [your] nose... surely this points even more strongly to the separation
> of the two systems in this poor overburdened head.

Of course you must realize that in changing the rules from the TT to the TTT (Total Turing Test, requiring not just linguistic, but robotic sensorimotor capacity) you've changed the game completely. For reasons I gave in "Minds, Machines and Searle," robotic function (e.g., transduction) is immune to Searle's argument. So not only can you not use it against him in court:

He couldn't even pull it off successfully. In fact, there's a REAL counterfactual for you: For whereas Searle could, both in practice and in principle, manipulate symbols without understanding them, he could not, even in principle, transduce optical input, at least without BEING the transducer. And if he really were the transducer, you would have the expert witnesses on your side this time, ready to testify that, even if he claimed he did not see, he could be exhibiting the clinical condition called blindsight (SOME system in there is "seeing").

> I imagine that you, Stevan, might find these objections quite
> congenial... some of the symbols in the Chinese Searle must be grounded
> in his physiology... proponents of the symbolic approach to AI and the
> systems reply to Searle would not disagree... But notice that the
> example needs to ASSUME this in order to be a convincing argument FOR
> it, although this assumption is usually not brought out into the open,
> and that it is the step from the Room to Searle-having-memorised-the
> -code, the standard reply to the systems reply, which makes it necessary.

You're quite right that I find the call for grounding congenial, but note that it's the switch to the TTT (a move I of course advocate) that necessitates it, NOT the speed/capacity/complexity factor you mentioned. And the TTT is already immune to Searle, so it's irrelevant to the court case or the Chinese Room Argument, which concerns only the TT. And as I've said before, what symbolic AI means by "grounding" is just connecting an independent symbol crunching module (which does virtually all of the work) to independent transducer/effector modules and hence the outside world "in the right way," whereas I think most of the hard work of connecting with the outside world "in the right way" will be done by nonsymbolic rather than symbolic function, and that a grounded system will not have an independent symbolic module: It will be hybrid nonsymbolic/symbolic through and through.

But even if we take AI's ecumenism about grounding at face value, it's enough foils the original thesis of Strong AI and explains why neither Chinese-Searle nor any other pure symbol-cruncher is understanding: Because understanding only occurs in grounded systems, and pure symbol crunching is ungrounded; it requires AT LEAST symbol-crunching-plus -transduction to understand. (Now there's a "systems reply" I could endorse!)

> You accuse the systems repliers of taking out a greater and greater
> counterfactual loan, but let me turn that accusation back at you (and
> Searle): the whole silly discussion, taken as an objection to the
> symbolic hypothesis, relies on counterfactuals which are so
> outrageous... [that] to draw serious methodological or scientific
> conclusions from this stuff is - putting it politely - to misplace ones
> confidence. Pat Hayes

And the interesting thing is that, despite the counterfactuals on both sides, serious methodological conclusions CAN be drawn: Pure symbol crunching will never get you to understanding.

Stevan Harnad

From: jc@cs.utexas.edu James Crawford

> I find this distinction between explicit and implicit rules
> somewhat slipery. Consider an expert system written in OPS5.
> Clearly the rules are represented explicitly. Now suppose I
> compile the rules into a parallel lisp, then to assembly code,
> and then to VLSI. It would be very hard to examine the result
> of this process and detect explicit rules. However, one could
> examine the switching mechanism inside a thermostat and make
> a somewhat valid argument that it implements a rule about when
> to switch on the heater.
>
> James Crawford

The definition says nothing about whether it should be hard or easy to do the semantic interpretation, just that a systematic interpretation must be possible. Compiled languages are easier (because they're closer to English) than machine code, but nothing essential rides on that. The switching mechanism in a thermostat, however, though "interpretable" in isolation, is certainly not amenable to combinatory semantics; it lacks precisely the kinds of systematic formal properties that make machine code so hard to interpret, yet systematically interpretable nonetheless. Now, what does seem an open question is what counts as a minimal symbol system, since we clearly have the full power of natural and artificial languages on one end and an inert rock at the other. How much combinatoriness does it take to have a semantic system?

Stevan Harnad

From: kube%cs@ucsd.edu (Paul Kube)

> (2) manipulated on the basis of EXPLICIT RULES that are
>
> (3) likewise physical tokens and STRINGS of tokens. The rule-governed
> symbol-token manipulation is based
>
> This excludes some implementations of Turing machines, and so is too
> strong. Of course, a TM must have a state table, but needn't operate by
> consulting a representation of it. Robert Cummins had a paper a few
> years back in which he described a TM whose states are implemented
> "directly" as states of a chemical system; what's wrong with that?
>
> Or to put it another way: What's a principled reason for requiring
> that the rules for manipulating symbols be explicitly represented,
> while not requiring the same of the rules for applying those rules?
> This regress has to stop somewhere, but I don't see that stopping it
> one place rather than another makes any difference for whether or not
> the system is symbolic. -- Paul Kube@ucsd.edu

You may be right that it's too strong, but perhaps you should propose a viable weakening of it that still distinguishes symbol systems from whatever their complement is. Anything that excludes Turing Machines -- which are, after all, the paradigm for all this -- is of course undesirable. As far as I can tell, the reason you need explicitness is in order to support systematicity: If the semantics is combinatory, the syntax must be too.

Stevan Harnad

----

mcdermott-drew@CS.YALE.EDU (Drew McDermott) of Yale University Computer Science Dept asked:

> Why is it necessary that a symbol system have a semantics in order to
> be a symbol system? I mean, you can define it any way you like, but
> then most AI programs wouldn't be symbol systems in your sense.
>
> Perhaps you have in mind that a system couldn't really think, or
> couldn't really refer to the outside world without all of its symbols
> being part of some seamless Tarskian framework... I think you have to
> buy several extra premises about the potency of knowledge
> representation to believe that formal semantics is that crucial.

I'd rather not define it any way I like. I'd rather pin people down on a definition that won't keep slipping away and thereby reducing all disagrements about what symbol systems can and can't do to mere matters of interpretation.

I gave semantic interpretability as a criterion, because it really seems to be one of the properties people have in mind when they single out symbol systems. However, semantic interpretability is not the same as having an intrinsic semantics, in the sense that mental processes do. But I made no reference to anything mental ("thinking," reference," "knowledge") in the definition.

So the only thing at issue is whether a symbol system is required to be semantically interpretable. Are you really saying that most AI programs are not? I.e., that if asked what this or that piece of code means or does, the programmer would reply: "Beats me! It's just crunching a bunch of meaningless and uninterpretable symbols."

No, I still think an obvious sine qua non of both the formal symbol systems of mathematics and the computer programs of computer science and AI is that they ARE semantically interpretable.

Stevan Harnad

From: mclennan%MACLENNAN.CS.UTK.EDU@cs.utk.edu

Steve Harnad has invited rival definitions of the notion of a symbol system. I formulated the following (tentative) definition as a basis for discussion in a connectionism course I taught last year. After stating the definition I'll discuss some of the ways it differs from Harnad's.

PROPERTIES OF DISCRETE SYMBOL SYSTEMS

A. Tokens and Types

1. TOKENS can be unerringly separated from the background.

2. Tokens can be unambiguously classified as to TYPE.

3. There are a finite number of types.

B. Formulas and Schemata

1. Tokens can be put into relationships with one another.

2. A FORMULA is an assemblage of interrelated tokens.

3. Formulas comprise a finite number of tokens.

4. Every formula results from a computation (see below) starting from a given token.

5. A SCHEMA is a class of relationships among tokens that depends only on the types of those tokens.

6. It can be unerringly determined whether a formula belongs to a given schema.

C. Rules

1. Rules describe ANALYSIS and SYNTHESIS.

2. Analysis determines if a formula belongs to a given schema.

3. Synthesis constructs a formula belonging to a given schema.

4. It can be unerringly determined whether a rule applies to a given formula, and what schema will result from applying that rule to that formula.

5. A computational process is described by a finite set of rules.

D. Computation

1. A COMPUTATION is the successive application of the rules to a given initial formula.

2. A computation comprises a finite number of rule appli- cations.

COMPARISON WITH HARNAD’S DEFINITION

1. Note that my terminology is a little different from Steve’s: his "atomic tokens" are my "tokens", his "composite tokens" are my "formulas". He refers to the "shape" of tokens, whereas I distinguish the "type" of an (atomic) token from the "schema" of a formula (composite token).

2. So far as I can see, Steve's definition does not include anything corresponding to my A.1, A.2, B.6 and C.4. There are all "exactness" properties -- central, although rarely stated, assumptions in the theory of formal systems. For example, A.1 and A.2 say that we (or a Turing machine) can tell when we're looking at a symbol, where it begins and ends, and what it is. It is important to state these assumptions, because they need not hold in real-life pattern identification, which is imperfect and inherently fuzzy. One reason connectionism is important is that by questioning these assumptions it makes them salient.

3. Steve's (3) and (7), which require formulas to be LINEAR arrangements of tokens, are too restrictive. There is noth- ing about syntactic arrangement that requires it to be linear (think of the schemata used in long division). Indeed, the relationship between the constituent symbols need not even be spatial (e.g., they could be "arranged" in the frequency domain, e.g., a chord is a formula comprising note tokens). This is the reason my B.5 specified only "relationships" (perhaps I should have said "physical rela- tionships").

4. Steve nowhere requires his systems to be finite (although it could be argued that this is a consequence of their being PHYSICAL systems). I think finiteness is essential. The theory of computation grew out of Hilbert's finitary approach to the foundations of mathematics, and you don't get the standard theory of computation if infinite formulas, rules, sets of rules, etc. are allowed. Hence my A.3, B.3, C.5, D.2.

5. Steve requires symbol systems to be semantically interpret- able (8), but I think this is an empty requirement. Every symbol system is interpretable -- if only as itself (essen- tially the Herbrand interpretation). Also, mathematicians routinely manipulate formulas (e.g., involving differen- tials) that have no interpretation (in standard mathematics, and ignoring "trivial" Herbrand-like interpretations).

6. Steve's (1) specifies a SET of formulas (physical tokens), but places no restrictions on that set. I'm concerned that this may permit uncountable or highly irregular sets of for- mulas (e.g., all the uncomputable real numbers). I tried to avoid this problem by requiring the formulas to be generatable by a finite computational process. This seems to hold for all the symbol systems discussed in the literature; in fact the formation rules are usually just a context-free grammar. My B.4 says, in effect, that there is a generative grammar (not necessarily context free) for the formulas, in fact, that the set of formulas is recursively enumerable.

7. My definition does not directly require a rule itself to be expressible as a formula (nearly Steve's 3), but I believe I can derive this from my C.1, C.2, C.3, although I wouldn't want to swear to it. (Here's the idea: C.2 and C.3 imply that analysis and synthesis can be unambiguously described by formulas that are exemplars of those schemata. Hence, by C.1, every rule can be described by two examplars, which are formulas.)

Let me stress that the above definition is not final. Please punch holes in it!

Bruce MacLennan Department of Computer Science The University of Tennessee

----

I think I like most of this, except the part about the semantic interpretability.

Stevan Harnad

From: miken@ai.mit.edu (Michael N. Nitabach) MIT AI Lab, Cambridge, MA

This is a very clear description of the definition of a symbolic system. I agree that all these elements are necessary; remove any one of them, and what is left is no longer a natural kind. It is a fascinating question, and one I would enjoy discussing in the mailing list, What serves as evidence that a system is explicitly, rather than implicitly, following a rule?

--Mike Nitabach

From: Stevan Harnad

The evidence would come from decomposability and recombinability while sustaining a systematic semantic interpretation. E.g., if there's an "If A then B" rule explicitly represented, you should be able to find the "If," the "A" and the "B, in other ruleful combinations, such as "If -A then C," etc.

Stevan Harnad

From: Beth Preston EFP@vms.cis.pitt.edu

At the risk of blowing my cover and sounding like the classically trained philosopher I am, I must take issue with your attempt to restrict the use of the term 'symbol' to its use in theory of computation. In the first place, the term was not invented by Church or Turing; it's been around since the ancient Greeks, and there have been theories of symbols and symbol use for the last two thousand years. Theory of computation is a very latecomer on this scene, and has no right to exclusive use of this term as far as I can see, although they are welcome to stipulate a technical use of it within their own domain if they like. In the second place, if you look in a dictionary you will find that the meaning given there for the term 'symbol' is usually something along the lines of 'something that stands for or represents something else' --a meaning very close to Pat Hayes suggestion that a symbol in the most general sense is a bearer of information. In short, I am very unwilling to let you just have this term for your own purposes since it is a term with a long philosophical history as well as an established career in common parlance.

In addition, I think your proposed restriction has some unfortunate consequences. One is with regard to the term 'representation'--a term with about the same sort of history and flexibility as 'symbol'. It seems intuitively right to say that a representation is a symbol. But if connectionist systems are non-symbolic, then it would follow that they are non-representational as well, a position which noone seems to hold. Which means that by restricting the term 'symbol' in the way you suggest, you will prevent people from making sense of the term 'representation' as well. In particular, you will prevent them from making sense of the intuition that the kinds of systematic/compositional symbol systems you are talking about contain representations, and that these representations have something to do with the things commonly referred to as representations in other sorts of systems (e.g., connectionist ones).

Which leads to the second unfortunate consequence of your position, namely that by dividing the universe up into the symbolic and the non-symbolic, and then restricting the sense of 'symbolic' so narrowly, you may effectively prevent people from noticing and making use of the many ways in which symbolic and non-symbolic (in your sense now) systems are similar. There are lots of things which have some of the properties of your symbol systems, but not others, and it would seem counterproductive if not an outright distortion to fail to take this into account in any analysis of the system's behavior. So again, I agree with Pat Hayes that your restrictive definition is likely to prove scientifically unfruitful--and I would add philosophically stultifying.

Beth Preston

From: Stevan Harnad

It seems to me that a long tradition of vague usage is not a reason against trying to be more precise and explicit. It's always good to try to come to some common agreement about what we mean when we use a term.

And my definition does not say anything about either representation or connectionism. I'm content to let the chips fall where they may there. I'd just like to see people sign off on what they are referring to when they speak of what is and is not, or can or cannot be symbolic. Certainly the position that everything (every structure, process, state) is symbolic (or that everything that is semantically interpretable is symbolic, which comes to the same thing) is not satisfactory because it simply deprives the category of any content.

Stevan Harnad

From: Dave.Touretzky@B.GP.CS.CMU.EDU

Okay, I'll take a shot at responding to Stevan's query. Of the eight criteria he listed, I take exception to numbers 2 and 3, that rules must be EXPLICIT and expressed as STRINGS of tokens.

In my recent work on phonology with Deirdre Wheeler (see "A connectionist implementation of cognitive phonology", tech report number CMU-CS-89-144), we define an architecture for manipulating sequences of phonemes. This architecture supports a small number of primitive operations like insertion, mutation, and deletion of phonemes. I claim that the rules for deciding WHEN to apply these primitives do not need to be represented explicitly or have a symbolic representation, in order for us to have a symbol system. It suffices that the rules' actions be combinations of the primitive actions our architecture provides. This is what distinguishes our phonological model from the Rumelhart and McClelland verb learning model. In their model, rules have no explicit representation, but in addition, rules operate directly on the phoneme sequence in a totally unconstrained way, mapping activation patterns to activation patterns; there are no primitive symbolic operations. Therefore their model is non-symbolic, as they themselves point out.

I also think the definition of symbol system Stevan describes is likely to prove so constrained that it rules out human cognition. This sort of symbol system appears to operate only by disassembling and recombining discrete structures according to explicit axioms. What about more implicit, continuous kinds of computation, like using spreading activation to access semantically related concepts in a net? How far the activation spreads depends on a number of things, like the branching factor of the semantic net, the weights on the links, and the amount of cognitive resources (what Charniak calls "zorch") available at the moment. (People reason differently when trying to do other things at the same time, as opposed to when they're relaxed and able to concentrate on a single task.)

Of course, spreading activation can be SIMULATED on a symbol processing system, such as a Turing machine or digital computer, but this raises the very important issue of levels of representation. What the Physical Symbol System Hypothesis requires is that the primitive atomic symbol tokens have meaning IN THE DOMAIN of discourse we're modeling. Although a Turing machine can be made to simulate continuous computations at any level of precision desired, it can only do so by using its primitive atomic symbols in ways that have nothing to do with the semantic net it's trying to simulate. Instead its symbols are used to represent things like the individual bits in some floating point number. To play the Physical Symbol Systems game correctly in the semantic net case, you have to choose primitives corresponding to nodes and links. But in that case there doesn't seem to be room for continuous, non-compositional sorts of computations.

Another problem I see with this definition of symbol system is that it doesn't say what it means in #5 to "rulefully combine" symbols. What about stochastic systems, like Boltzmann machines? They don't follow deterministic rules, but they do obey statistical ones. What about a multilayer perceptron, which could be described as one giant rule for mapping input patterns to output patterns?

Dave Touretzky

From: Stevan Harnad

> the rules for deciding WHEN to apply these primitives do not need to
> be represented explicitly or have a symbolic representation, in order
> for us to have a symbol system.

Fine, but then what do YOU mean by "symbol system." You seem to be presupposing that whatever your phonological net is, it MUST be a symbol system. It is not enough to reject my criteria just because they may reject your net. You must provide rival criteria of your own, that include not only your net, but what people want to pick out with the term "symbol system," and that exclude what they want to exclude.

> I also think the definition of symbol system Stevan describes is
> likely to prove so constrained that it rules out human cognition.

That's only a problem if it's true that cognition is just symbol manipulation! I wasn't trying to define or even include cognition, particularly. I was just trying to settle on a definition of "symbol system." (There's an awful lot of circularity on this topic...)

> Of course, spreading activation can be SIMULATED on a symbol processing
> system, such as a Turing machine or digital computer, but this raises
> the very important issue of levels of representation... the primitive
> atomic symbol tokens [must] have meaning IN THE DOMAIN of discourse
> we're modeling.

I'm trying to avoid the hermeneutic hall of mirrors created by projecting interpretations onto symbol systems. This means keeping it clear in our minds that, apart from a symbol systems's having to be systematically INTERPRETABLE, the interpretations themselves (and hence their level, and their "domain of discourse") are distinct from the symbol system, which is in reality merely uninterpreted syntax. (See my reply to Crawford above about programming languages and virtual machines.)

> doesn't say what it means in #5 to "rulefully combine" symbols. What
> about stochastic systems... multilayer perceptrons...

What about them indeed?...

Stevan Harnad

To: Noam Chomsky

Dear Noam:

I've re-entered your letter electronically because although, at your request, I'm not circulating it for discussion to the Symbol Grounding Discussion Group, skywriting has given me the happy habit of requoting extensively in focusing my replies. (If you give me permission, though, I'd like to branch a copy of my reply to John Searle, who has an obvious interest in this topic, and to Ned Block, who is hosting two talks by me on this topic on January 23rd at MIT, to which I hope you'll come!)

You wrote:

> Dear Stevan, Thanks for sending me your interesting and thoughtful
> papers. I think they add a lot to the discussion, but they don't clear
> up my own confusion about what the fuss is about. This is not intended
> as a contribution to the ongoing discussion, but rather a request for
> clarification. Let me sketch a question which is based on some [points]
> raised by Pylyshyn and Stich about the whole business.

I will try to show point for point what the fuss is about. Unless I have misunderstood the point in your letter, I think I can see what the basis of the confusion is.

> We're concerned with a device that can pass the TT, receiving a
> sentence as input and giving an appropriate response as output, as,
> say, Searle would do it in English, and thus "fooling" the
> experimenter.

Here's the first potential point of confusion: We are not concerned with a device that "fools" the experimenter (i.e., acts as if it understands, but doesn't really understand). If you accept that the device is just fooling then you've already conceded the conclusion of Searle's argument, which is that such a device -- and the kind of device does matter, as we shall see -- a computer, a mere symbol manipulator, cannot really understand!

The thesis of "Strong AI" -- the one Searle is trying to show, with his thought experiment, to be false -- is that the computer REALLY understands, just as you or I do; not that it fools the experimenter. And the punchline will be that, whatever function "real understanding," like ours, consists of, it is not just symbol manipulation, which is the only thing a computer can do.

(For the record, I happen to doubt, for reasons I have written about, that a pure symbol manipulator could actually pass the TT in practice -- i.e., fool the experimenter for a lifetime -- even if, according to one interpretation of the Church-Turing Thesis, this should be possible in principle. Note, though, that this is only a performance doubt. We will accept this premise arguendo, even if it is counterfactual, to see whether it is compatible with the conditional conclusion that IF a pure symbol-manipulator could pass the TT, THEN it would understand.)

> Suppose that, to a first approximation, Searle's performing this task
> works like this: External stimuli (auditory, visual) are mapped into
> some systems of his mind/brain by the mapping M(e) [e = english], which
> yields, as "output," representations (states, whatever one likes) that
> incorporate whatever the language has to say that is semantically
> relevant (say, for concreteness, LF-representation [logical form],
> though it doesn't matter here what one chooses).

Let me make sure I understand what is being supposed here: I take it that the M(e) mapping is just from the sound and sight of words (not the objects they stand for) onto internal versions of those words, in the form of symbol strings in some code, realized in some hardware, right? So far, so good, since we're still just talking about symbol tokens.

I also assume that by "semantically relevant" you just mean the same thing you would mean about a passage in a textbook: It contains symbol strings that are semantically interpretable (and relevant). You are not, I hope, presupposing that the semantic interpretation that our minds apply to those marks on the paper in understanding them is somehow intrinsic to the paper rather than derived from our minds; for that too is one of the points at issue.

One of the ways to put it is that both the Chinese Room Argument and the Symbol Grounding Problem show that there is a great difference between (1) systems whose symbols are merely semantically interpretable, like a book, and (2) ("grounded") systems, like our brains, that somehow include the semantic interpretations of their symbols, as our brains do. Your clause about "incorporating everything the language has to say that is semantically relevant" is ambiguous in this regard. I will assume you mean only semantically interpretable, like a book; otherwise you would be begging the question.

> Then some other component of his mind/brain H [homunculus], perhaps
> accessing other kinds of information, selects an approprate response,
> formulating it with an LF-representation, then mapping it by M(e)' to
> an arbitrary output -- passing the test.

Here too, we have to be very explicit about the kind of thing you are imagining "H" to be doing. For Strong AI, it must be only symbol manipulation -- manipulating symbol tokens on the basis of purely syntactic rules that operate on the tokens' (arbitrary) shapes, not their meanings (which have not yet been grounded in any way). No one would doubt that SOME internal function in our brains (not necessarily best viewed as homuncular, by the way) operates on the words we hear or see in such a way as to turn them into the kinds of interpreted thoughts we understand, but what's at issue here is the KIND of function that might be.

Searle's objective is to show that it can't be just symbol manipulation. If you agree -- if you think that the structures and processes involved in "accessing other kinds of information," the structures and processes that Searle would have to BE and DO in order to pass the TT, would not be just symbol manipulation -- then again you've already agreed with Searle. But if you instead believe that what "H" does could be just symbol manipulation, and that all this collective symbol manipulation alone, when implemented, would understand, then let's go on:

> Suppose now that Wang, a speaker of Chinese, happens to be the same as
> Searle in the relevant respects, except that instead of M(e) and M(e)',
> he has M(c) and M(c)' [c = chinese]. This wired-up Searle now passes
> the TT for Chinese, just as he did for English.

Again, you have to be very careful not to beg the question. The question is whether a pure symbol manipulator that passes the (Chinese) TT is really understanding Chinese. No one doubts (modulo the other-minds problem) that a real Chinese speaker understands Chinese.

And we are not "wiring" anything. Of course if we "rewired" Searle's brain we could make him understand Chinese. But that wasn't the question. It was whether a pure symbol manipulator that passed the TT would understand Chinese.

To show that it wouldn't, Searle is merely showing that he can DO everything the symbol manipulator can DO -- perform all of its functions -- yet still not understand Chinese. He can even memorize all the symbols and not only DO everything it does, but BE everything it is, and still not understand Chinese.

Note that any OTHER internal functions (say, nonsymbolic ones, like transduction and analog processes) that a real Chinese speaker's brain (and Searle's brain) may have that a computer lacks are completely irrelevant in this contest. The only thing that's relevant is what the computer can do (i.e., symbol manipulation): if Searle can do all that without understanding, then he has made it look pretty arbitrary to continue insisting that, nevertheless, the computer DOES understand when it's doing exactly the same thing!

> Does wired-up Searle understand Chinese? One of two answers is
> possible. Either he does, in which case there is no problem raised by
> the Chinese room argument; or he doesn't, in which case we are left
> with the amazing and highly counter-intuitive conclusion that the
> understanding of Chinese depends on being able to carry out the
> mapping algorithms for Chinese. Surely no one would accept that. So,
> wired-up Searle understands Chinese. In fact, he's essentially
> identical with Wang.

As I’ve said, "wiring up" has nothing to do with it. Searle merely does what the computer does: he manipulates meaningless symbols. From my knowledge of neuropsychology and from my common sense experience there is nothing at all that suggests that memorizing and manipulating a bunch of meaningless symbols can "rewire" a brain in any way that is nearly radical enough to give rise to an understanding of a foreign language. That’s certainly not the way we actually learn foreign languages (as I pointed out in the first [easier] version of my "Chinese-Chinese Dictionary-Go-Round" in The Symbol Grounding Problem).

(By the way, if your mapping "M" had been from objects, actions and states of affairs in the world to symbolic representations (and vice versa) rather than just from verbal inputs to symbolic representations (and vice versa), then, according to my symbol grounding theory at least, a version of the "highly counter-intuitive conclusion that understanding depends on being able to carry out the mapping algorithms [i.e., sensorimotor grounding]" would in fact turn out to be quite correct!)

> Does wired-up Searle’s brain understand Chinese? Of course not, any
> more than Searle’s brain understands English. For reasons irrelevant
> to cognitive science, in normal usage we speak of people understanding
> things, not their brains. No problem here, and no interest in the
> fact. Clearly, cognitive science is no more interested in preserving
> features of ordinary usage than physics is. Have we achieved anything
> by the simulation, yielding wired-up Searle? Of course not; wired-up
> Searle is an awful model, because it teaches us nothing. Same with the
> Chinese room.

I hope that by now it is clear that "wired-up Searle" is not at issue, only the possibilities of pure symbol manipulation. And that in the memorized-symbols version of the Chinese Room, there is no superordinate "system" at all to point to, other than Searle himself. So considerations of ordinary usage have nothing to do with it.

As to what it teaches us: I guess it’s time for me to trot out old J.H. Hexter again:

"in an academic generation a little overaddicted to "politesse," it may be worth saying that violent destruction is not necessarily worthless and futile. Even though it leaves doubt about the right road for London, it helps if someone rips up, however violently, a ‘To London’ sign on the Dover cliffs pointing south..."

> So what is the discussion all about? -- Best, Noam

It is about the fact that, although Searle doesn’t know the right road to understanding, he’s shown that you’ll never get there via pure symbol manipulation. (In my own papers I’ve suggested some nonsymbolic alternative routes, how they are immune to Searle, and how and why they might work.)

Best, Stevan

From: miken@ai.mit.edu (Michael N. Nitabach)

Although I am very wary of engaging in any discourse regarding the Chinese Room argument, since it seems impossible to say something which hasn't been said countless times before, I will give it a try. My feeling is that, to a certain extent, the Chinese Room is arguing against a view of cognition that is not really held by any serious cognitive scientist, even if they are rightly placed in the "symbolist" camp.

First, even as staunch symbolists as Fodor and Pylyshyn would (and have done so) agree that an essential element of a cognizing system that exists in the real world is a set of transducers and effectors which translate to and from (respectively) symbolic representations of and analogues of events in the physical world.

Second, to the extent that contemporary accounts of Representational Theory of Mind (RTM) derive from the Husserlian heritage, they include a notion of the Gegebenheitsweise (mode of givenness) of a conscious act. Viewed simplistically, the Gegebenheitsweise encodes the species of act through which a symbolic representation presents itself to mind. So, by current accounts, "belief", "desire", "hope", "expectation", etc. are the Gegebenheitsweise of the acts of "believing that P", "hoping that P", etc., where P, the object of the act (Husserl's noema), is a symbolic representation which can be identical in each of these different species of act. In contemporary terms, "to believe that P" is to be in a certain functional relation to the symbolic representation, P, while "to desire that P", is to be in some other particular functional relation to the same symbolic representation P. So, on this account, to be cognizant of ("to understand") some domain of knowledge is to be in some (perhaps potential) functional relation to a symbolic representation of that knowledge.

It is thus completely consistent with RTMs, symbolist theories of mind, to insist that neither Searle (nor any extended system in the Chinese Room) understands Chinese; in no way has the correct type of functional relation (Gegebenheitsweise) to the Chinese characters been instantiated which would justify the attribution of understanding Chinese.

To summarize, current RTMs require the presence of at least two elements to justify attributions of understanding: (1) a symbolic representation of the object of understanding and (2) an appropriate species of functional relation between the understanding entity and this representation. This analysis has crucial methodological implications for AI and for all of the cognitive sciences. Complete accounts (or artificial implementations) of cognition must include explanations (or instantiations) of both the symbolic representations which are the objects of mental acts, *and* the appropriate functional relations which cognizers bear to those representations.

The Chinese Room *is* an example of a system which doesn't understand, but *not* due to a fundamental flaw in all symbolist theories of mind. This system doesn't understand because it fails to embody at least one of the elements of a class of symbolist theories of mind: RTMs. My feeling is that, at least in AI, analysis of the nature of representations has been overemphasized, while analysis of "modes of givenness" (functional relations) has been given short shrift. Perhaps this reflects what I see as a great disparity in the difficulty of these two tasks.

[I apologize in advance if this is just a rehash of previously stated views.]

Mike Nitabach

Reply from: Stevan Harnad

> even staunch symbolists... agree that an essential element of a
> cognizing system that exists in the real world is a set of transducers
> and effectors which translate to and from (respectively) symbolic
> representations of and analogues of events in the physical world.

Yes, this has been said before. The answer is: If this means that in order to understand Chinese it is NOT enough to be just a symbol system that passes the (linguistic) Turing Test (TT), then this is simply to agree with Searle!

(And I have argued that the problem of "connecting" transducers & effectors to a symbol system in "the right way" -- a way that will successfully "translate" to and from symbols, analogs and the world -- will be no simple matter, and that most of the real work may turn out to be nonsymbolic. This is the Symbol Grounding Problem, and it is not at all clear what is left of the symbolic theory of mind once it is confronted.)

> It is thus completely consistent with RTMs, symbolist theories of mind,
> to insist that neither Searle (nor any extended system in the Chinese
> Room) understands Chinese; in no way has the correct type of functional
> relation (Gegebenheitsweise) to the Chinese characters been
> instantiated which would justify the attribution of understanding
> Chinese... Complete accounts (or artificial implementations) of
> cognition must include explanations (or instantiations) of both the
> symbolic representations which are the objects of mental acts, *and*
> the appropriate functional relations which cognizers bear to those
> representations.

Again, to the extent that those functional relations are nonsymbolic and essential to having a mind, Searle’s conclusion is corroborated. Searle was only trying to refute the notion that understanding (thinking, being intelligent, having a mind) IS merely symbol manipulation, and hence that any system (whether the computer or Searle) that passes the TT must understand.

If you call for more than symbol manipulation in order to be a system that understands -- either more internal functions (e.g., transducer/effector function) or more performance capacity (e.g., the ability to pass the Total Turing Test, which includes all of our robotic interactions with objects in the world) -- you are simply agreeing with me.

> The Chinese Room *is* an example of a system which doesn’t understand,
> but *not* due to a fundamental flaw in all symbolist theories of mind.
> This system doesn’t understand because it fails to embody at least one
> of the elements of a class of symbolist theories of mind: RTMs. My
> feeling is that, at least in AI, analysis of the nature of
> representations has been overemphasized, while analysis of "modes of
> givenness" (functional relations) has been given short shrift.

I have no idea what "modes of givenness" or their analysis might amount to. But I have tried to show that pure symbol systems are ungrounded. And that sounds like "a fundamental flaw in all symbolist theories of mind" to me, at least those (like Strong AI and the Systems Reply you've heard so much about in this discussion) that insist that the Chinese Room DOES understand.

> the Chinese Room is arguing against a view of cognition that is not
> really held by any serious cognitive scientist, even if they are
> rightly placed in the "symbolist" camp.

I guess you either haven't been following this discussion or you don't consider any of my opponents serious; but let me assure you that plenty of serious people do hold the view in question (and that it won't be the first time that serious people have been wrong...).

Stevan Harnad

From: cam@aipna.ed.ac.uk (Chris Malcolm) Organization: Dept of AI, Edinburgh University, UK.

In your original posting you (Stevan Harnad) said:

So the mere fact that a behavior is "interpretable" as ruleful does not mean that it is really governed by a symbolic rule. Semantic interpretability must be coupled with explicit representation (2), syntactic manipulability (4), and systematicity (8) in order to be symbolic. There is a can of worms luring under that little word "coupled"! What I take it to mean is that this symbolic rule must cause the behaviour which we interpret as being governed by the rule we interpret the symbolic rule as meaning. Unravelled, that may seem stupendously tautologous, but meditation on the problems of symbol grounding can induce profound uncertainty about the status of supposedly rule-governed AI systems. One source of difficulty is the difference between the meaning of the symbolic rule to the system (as defined by its use of the rule) and the meaning we are tempted to ascribe to it because we recognise the meaning of the variable names, the logical structure, etc.

Brian Smith's Knowledge Representation Hypothesis contains a nice expression of this problem of "coupling" interpretation and causal effect, in clauses a) and b) below.

Any mechanically embodied intelligent process will be be comprised of structural ingredients that a) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and b) independent of such external semantical attribution, play a formal but causal and essential role in engendering the behaviour that manifests that knowledge.

[Brian C. Smith, Prologue to "Reflection and Semantics in a Procedural Language" in "Readings in Knowledge Representation" eds Brachman & Levesque, Morgan Kaufmann, 1985.] It is not at all clear to me that finding a piece of source code in the controlling computer which reads IF STRING_PULLED THEN DROP_HAMMER is not just a conjuring trick where I am misled into equating the English language meaning of the rule with its function within the computer system [Drew McDermott, Artificial Intelligence meets Natural Stupidity, ACM SIGART Newsletter 57, April 1976]. In simple cases with a few rules and behaviour which can easily be exhaustively itemised we can satisfy ourselves that our interpretation of the rule does indeed equate with its causal role in the system. Where there are many rules, and the rule interpreter is complex (e.g. having a bucketful of ad-hoc conflict-resolution prioritising schemes designed to avoid "silly" behaviour which would otherwise result from the rules) then the equation is not so clear. The best we can say is that our interpretation is _similar_ to the function of the rule in the system. How reliably can we make this judgment of similarity? And how close must be the similarity to justify our labelling an example as an instance of behaviour governed by an explicit rule?

Why should we bother with being able to interpret the system's "rule" as a rule meaningful to us? Perhaps we need a weaker category, where we identify the whole caboodle as a rule-based system, but don't necessarily need to be able to interpret the individual rules. But how can we do this weakening, without letting in such disturbingly ambiguous exemplars as neural nets?

Chris Malcolm

Reply to Malcolm From: Stevan Harnad

> What I take [you] to mean is that [the] symbolic rule must cause the
> behaviour which we interpret as being governed by the rule we interpret
> the symbolic rule as meaning... meditation on the problems of symbol
> grounding can induce profound uncertainty about the status of
> supposedly rule-governed AI systems. One source of difficulty is the
> difference between the meaning of the symbolic rule to the system (as
> defined by its use of the rule) and the meaning we are tempted to
> ascribe to it because we recognise the meaning of the variable names,
> the logical structure, etc.

I endorse this kind of scepticism -- which amounts to recognizing the symbol grounding problem -- but it is getting ahead of the game. My definition was only intended to define "symbol system," not to capture cognition or meaning.

You are also using "behaviour" equivocally: It can mean the operations of the system on the world or the operations of the system on its symbol tokens. My definition of symbol system draws only on the latter (i.e., syntax); the former is the grounding problem.

It is important to note that the only thing my definition requires is that symbols and symbol manipulations be AMENABLE to a systematic semantic interpretation. It is premature (and as I said, another problem altogether) to require that the interpretation be grounded in the system and its relation to the world, rather than just mediated by our own minds, in the way we interpret the symbols in a book. All we are trying to do is define "symbol system" here; until we first commit ourselves on the question of what is and is not one, we cannot start to speak coherently about what its shortcomings might be!

(By the way, "meaning to us" is unproblematic, whereas "meaning to the system" is highly contentious, and again a manifestation of the symbol grounding problem, which is certainly no definitional matter!)

> It is not at all clear to me that finding a piece of source code in the
> controlling computer which reads IF STRING_PULLED THEN DROP_HAMMER is
> not just a conjuring trick... In simple cases with a few rules and
> behaviour which can easily be exhaustively itemised we can satisfy

> ourselves that our interpretation of the rule does indeed equate with
> its causal role in the system. Where there are many rules... The best
> we can say is that our interpretation is _similar_ to the function of
> the rule in the system. How reliably can we make this judgment of
> similarity? And how close must be the similarity to justify our
> labelling an example as an instance of behaviour governed by an
> explicit rule?

Again, you’re letting your skepticism get ahead of you. First let’s agree on whether something’s a symbol system at all, then let’s worry about whether or not its "meanings" are intrinsic. Systematic interpretability is largely a formal matter; intrinsic meaning is not. It is not a "conjuring trick" to claim that Peano’s system can be systematically interpreted as meaning what WE mean by, say, numbers and addition. It’s another question altogether whether the system ITSELF "means" numbers, addition, or anything at all: Do you see the distinction?.

(No one, has actually proposed the Peano system as a model of arithmetic understanding, of course; but in claiming, with confidence, that it is amenable to being systematically interpreted as what we mean by arithmetic, we are not using any "conjuring tricks" either. It is important to keep this distinction in mind. Number theorists need not be confused with mind-modelers.)

But, as long as you ask, the criterion for "similarity" that I have argued for in my own writings is the Total Turing Test (TTT), which, unlike the conventional Turing Test (TT) (which is equivocal in calling only for symbols in and symbols out) calls for our full robotic capacity in the world. A system that can only pass the TT may have a symbol grounding problem, but a system that passes the TTT (for a lifetime) is grounded in the world, and although it is not GUARANTEED to have subjective meaning (because of the other minds problem), it IS guaranteed to have intrinsic meaning.

(The "Total" is also intended to rule out spurious extrapolations from toy systems: These may be symbol systems, and even -- if robotic -- grounded ones, but, because they fail the TTT, there are still strong grounds for skepticism that they are sufficiently similar us in the relevant respects. Here I do agree that what is involved is, if not "conjuring," then certainly wild and unwarranted extrapolation to a hypothetical "scaling up," one that, in reality, would never be able to reach the TTT by simply doing "more of the same.")

> Why should we bother with being able to interpret the system’s "rule" as
> a rule meaningful to us?

Because that’s the part of how you tell whether you’re even dealing with a formal symbol system in the first place (on my definition).

Stevan Harnad

LEVELS OF INTERPRETATION AND THE EXPLICITNESS CRITERION

Paul Kube kube%cs@ucsd.edu wrote:

> Syntactic and semantic systematicity, whatever it is, is defined over
> the symbols that the system manipulates; roughly speaking, they have
> to constitute a language for the system. So far as I can see, this
> requires that the system manipulate them systematically, i.e. in a
> rule-described way, but doesn't require that the manipulation involve
> the consultation of explicitly represented, symbolic rules. Of course
> to be a symbol manipulator, it must really be manipulating explicit
> symbols, not just be describable as manipulating symbols; but again I
> don't see what difference it makes whether the systematicity in this
> manipulation is rule-governed or just rule-described, since it would
> be as systematic in either case.
>
> By the way, your position seems to get you in the Wittgensteinian
> regress with a vengeance; after all, what makes your explicit rules
> *symbolic*? Presumably, the fact that they're manipulated on the
> basis of further symbolic rules, etc.

Again, I think these are good points, and suggest ways in which the definition of "symbol system" may need to be sharpened. Are you suggesting that the six rules left if one stars 2 and 3 are sufficient to define a symbol system? I think the main reason that the rules need to be explicit rather than implicit is to ensure that THEIR interpretation has the requisite composite and combinatorial syntax to be systematic too -- in fact, that's all explicitness consists in, really. How/why should hard-wired or holistic "rules" decompose in a systematic way? Or are you suggesting that they needn't (or better not, on pain of regress)?

(1) a set of arbitrary PHYSICAL TOKENS (scratches on paper, holes on a tape, events in a digital computer, etc.) that are

*(2) manipulated on the basis of EXPLICIT RULES that are

*(3) likewise physical tokens and STRINGS of tokens. The rule-governed symbol-token manipulation is based

(4) purely on the SHAPE of the symbol tokens (not their "meaning"), i.e., it is purely SYNTACTIC, and consists of

(5) RULEFULLY COMBINING and recombining symbol tokens. There are

(6) primitive ATOMIC symbol tokens and

(7) COMPOSITE symbol-token strings. The entire system and all its parts -- the atomic tokens, the composite tokens, the syntactic manipulations (both actual and possible) and the rules -- are all

(8) SEMANTICALLY INTERPRETABLE: The syntax can be SYSTEMATICALLY assigned a meaning (e.g., as standing for objects, as describing states of affairs).

Stevan Harnad

From: EFP@vms.cis.pitt.edu Beth Preston

> about connectionism: it seems to me (and tell me if this doesn't seem
> right to you) that the systematicity and compositionality requirements
> in your definition exclude connectionist systems. But this is
> problematic, since connectionist systems are widely said to be
> computational systems, and since you say you want to tie the notion of
> symbol back to theory of computation. So now what do we say--that
> connectionist systems don't compute either? But then if we say that how
> do we make sense of the fact that these systems are built and
> maintained by trained practitioners of computation? and that they are
> built to do the same sorts of tasks that 'symbolic' systems do, e.g.,
> language processing? Or that they manipulate representations? (Given
> that 'symbol' and 'representation' are interchangeable terms in many
> contexts.) We don't have any basis for explaining these similarities
> if we draw the line between symbolic and non-symbolic where
> you want to draw it. Beth Preston

What's critical is that if a symbol system is interpreted as meaning or representing something, then that interpretation should be sustainable in a systematic way. Otherwise anything can be interpreted willy-nilly as meaning or representing anything. If a connectionist system can sustain the combinatorics of a systematic interpretation (say, this state means "The cat is on the mat"), fine, its a symbol system. Otherwise it's only a symbol system in the less interesting sense that the only systematic interpretation it can sustain is that of being a connectionist system (with units, layers, weights, activations, etc.)! Being symbolic in this sense is trivial, given that virtually all connectionist networks are actually symbolic simulations of connectionist networks.

Stevan Harnad

From: geller@vienna.njit.edu (james geller)

Last night I read your statement...

"Who knows what intentionality is? Who cares?"

(That's how far back I am in my mail...).

This statement made me think of the following question. Do you think that PEOPLE, e.g. philosophers, sometimes use terms that are insufficiently grounded?

Jim

From: harnad (Stevan Harnad)

Yes, I think people, including philosophers, sometimes use terms that are insufficiently grounded (or even downright incoherent or vacuous). I just think that ENOUGH of them have to use terms that are grounded ENOUGH to get by and make do.

Stevan

To: Franklin Boyle

> Responding to your request for life-signs, yes please KEEP me on the
> list. I read every message and hardcopy many. I happen to essentially
> agree with everything you've said. I guess I'm rather amazed at the
> ideas and responses of some of your critics.

Frank, thanks for the kind words. You ask:

> Are there many who generally agree with you, since most of what I read
> is in disagreement?

I haven't taken a poll, but I would say that the proportions IN THIS POPULATION (mainly AI) hover at about 50/50. Only a vocal minority actually spar with me (and I don't think they're the best thinkers around -- but I also don't think many of the best thinkers are on that side of the 50/50 fence either...

> When you discuss the two Rube Goldberg devices, you describe the first
> (the non-computer device) as the one with "pure causal throughput". I
> understand what you mean to say in using this phrase, but if somehow
> it's meant as a distinguishing characteristic, then at what point is
> the second device not purely causal?

Both devices are of course causal. The second, BESIDES being causal, also meets the criteria for being the implementation of a symbol system.

Stevan Harnad

From harnad Wed Nov 29 21:33:39 1989 To: searle@cogsci.berkeley.edu Subject: Noam's letter Cc: block@cogito.mit.edu, chomsky@cogito.mit.edu

To: John Searle

John,

> I could not understand Noam's argument. Can you send the whole
> communication?

Noam's complete letter was there, excerpted with my replies. To reconstruct it all you have to do is pull together all the pieces. Nothing was left out. But note that Noam allowed me to make personal use of the letter but NOT to post it for discussion, so I have not sent it, alas, to the Symbol Grounding Discussion Net, only to you and Ned Block. Note also that he has already replied to my reply (again NOT for posting to the Net) and that I'm about to reply to him again. You'll have a copy of that too.

Noam's argument is quite simple, but not successful, in my view. He wants to bracket the question of whether or not mental function is just symbol crunching, addressing only the question of whether we can learn anything at all from your Chinese Room Argument one way or the other. He wants to conclude that we do not, and that it is all just a misunderstanding related to our ordinary usage when we say who or what understands.

Noam suggests that all you could imitate would be the "chip" that gives symbolic inputs (in "Logical Form," say) to and takes symbolic outputs from the "homunculus" that does the real work, whatever that might be, and that therefore you haven't shown anything. The point he's missing, of course, is that the proponents of Symbolic AI were trying to model the homunculus itself, not just what gave to and took from it; that they wanted to do it all by symbol crunching and to declare the symbol cruncher itself to be really understanding; that you could do all of that and not understand, and that therefore you had shown that that kind of homunculus couldn't understand either.

Moral: You can't "bracket" the symbol-crunching issue, and it's NOT all just a big fuss about the ordinary language attribution of understanding to the person rather than the brain or a part of it.

I've added the email address you sent me to the Symbol Grounding List. I'm half way through your article and enjoying it. What about AAAI?

Cheers, Stevan

----

1. On Chomsky and Theoretical Inference vs. Semantic Interpretation (Nitabach; with Reply) 2. On Memorizing Symbols (Douthat) 3. On Explicit Rules (Dyer; with Reply) 4. On Working Hypotheses (McDermott; with Reply) 5. On Causal Grounding (Nitabach; with Reply)

From: miken@ai.mit.edu (Michael N. Nitabach)

Subject: Defining Symbol Systems

ME> What serves as evidence that a system is explicitly, ME> rather than implicitly, following a rule?

SH> The evidence would come from decomposability and recombinability while SH> sustaining a systematic semantic interpretation. E.g., if there's an SH> "If A then B" rule explicitly represented, you should be able to find SH> the "If," the "A" and the "B, in other ruleful combinations, such as SH> "If -A then C," etc.

I agree that this provides a reasonable *definition* for explicit, as opposed to implicit, rule following. However, what I am looking for is some methodological criterion or heuristic for *empirically* deciding whether some observed behavior is explicit or implicit rule-following. In cases where we have direct access to the innards of a system, one can simply apply your definition, e.g. to the analysis of a computer program. However, in the cases of greatest interest this is not always possible.

For example, consider natural language processing by human beings. Here, we clearly do not have direct access to the mechanistic innards of the language faculty. Thus, the only evidence we have that rules of grammar are explicitly followed in producing (or comprehending) sentences is something like a parsimony criterion; it is easy to imagine a device which explicitly uses a grammar in producing and interpreting sentences, yet it would require some hideously complex, and probably ad hoc, construc- tions to generate a device which just happens, due to its physical structure, to follow the rules of grammar. It is considerations of this type which people such as Chomsky take as legitimate evidence that grammar is an explicitly represented knowledge structure. Given the naturalness with which we explain grammaticality as an explicitly rule-governed phe- nomenon, and *given the absence of any alternate suggestion as to a possible implicit basis*, we are justified in concluding that the grammaticality of natural language *is* explicitly rule-governed. And, this conclusion is subject to future revision on the basis of new evidence, as are the results of any empirical science.

On the other hand, it is the fact that this is the *only* type of evidence available which leads people such as Hilary Putnam to deny that claims of explicitness grounded in this way are justified. Chomsky would reply that this is the only type of evidence we have in *any* empirical science (non-demonstrative inference to a provisional best explanation), and that it is unfair to subject the cognitive sciences to a more stringent criterion of empirical justification than that which is accepted in other sciences.

Let us apply this criterion to a simple example to see how it works. Consider two types of thermostats: one is based on a bimetallic strip attached to a switch, while the other has a temperature sensor connected to a microprocessor which itself is connected to a switch. If we were to only look at these devices as black boxes, we would be forced to the same conclusion in both cases; based on the "Chomskian criterion" we would be justified in concluding that both of these devices explicitly follow the rule "if temperature is below x degrees, then turn the switch on". Note, however, that if we look inside the boxes, we are then forced into different conclusions. When we see the bimetallic strip, we must conclude that this device is only implicitly following the rule just stated, and thus discard our earlier conclusion. When we see the microprocessor, and examine its program, we actually find an explicit representation of the rule, and thus retain our earlier conclusion.

The above example highlights the curiousness of our position with respect to attributions of reality to mental representations of rules. On a cynical view, we see attributions of explicitness as convenient fictions which work well in predicting the behavior of an organism, but which we re- tain only due to our ignorance of the actual inner workings of the mind/brain. Or, in some cases we might retain these fictions because they are simpler to use in predicting the behavior of a system than a "real" description of its mechanism. (This is Dennett's justification for adopting the "intentional stance" toward a system.) On the other hand, someone with a more realist bent would say that this analysis would apply equally well to the theoretical entities of *any* empirical science, and the only relevance of this type of cynicism is quite ordinary--namely, that the attributions of reality applied to the theoretical entities of any empirical science are provisional, and subject to refutation at any time by new observations. But, this is simply what makes a science empirical, i.e. distinct from math- ematics.

So, while your formal criterion for explicitness is very reasonable, rather than solving the problem, it actually provides a base on which a serious methodological controversy sits. My purpose in this posting is to (1) elicit comment on this "Chomskian criterion", and (2) to elicit alternate proposals for the empirical justification of attributions of explicitness to mental representations.

Mike Nitabach

Reply From: Stevan Harnad

SEMANTIC INTERPRETATION IS NOT THE SAME AS THEORETICAL INFERENCE

MIke Nitabach wrote:

> [We need] some methodological criterion or heuristic for *empirically*
> deciding whether some observed behavior is explicit or implicit
> rule-following... [where we] do not have direct access to the
> mechanistic innards...

I'm afraid this conflates two things: The explicitness of a symbolic rule, on the one hand, and the reality of a theoretically inferred entity on the other. Your proposed "Chomskian criterion" is likewise ambiguous in this respect.

Explicitness has been proposed as one of the criterial features for a symbol system. Its purpose, as far as I can tell, is to ensure that the system will support a systematic semantic interpretation. Explicit rules and representations, because decomposable and combinatory, CAN sustain such systematic interprations, whereas implicit, autonomous, holistic "rules" and "representations" cannot.

Explicitness is a formal criterion. It need not have anything to do with behavior, organisms, innards, or even empirical science. You can ask, for example, whether or not a particular axiom is explicitly represented in a formal axiomatic system in mathematics, or in a computing device. The only "empirical" criterion you need -- regardless of whether you see the system's "innards" or merely infer them -- is sufficient systematicity to sustain your semantic interpretation of the rules, representations, and their components and interrelations.

The reality of theoretical entities is another issue, and there of course I would agree with the "Chomskian criterion" (of data-fit, parsimony, and no better rivals). But the reality of theoretical entities and the systematicity of semantic interpretations are not the same thing.

> the only evidence we have that rules of grammar are explicitly followed
> in producing (or comprehending) sentences is something like a parsimony
> criterion... Chomsky take[s this] as legitimate evidence that grammar
> is an explicitly represented knowledge structure.... [i.e.] that
> [data-fit, parsimony and absence of a better rival theory] is the only

> type of evidence we have in *any* empirical science

You are closer to the source at Olympus than I am, but my recollection (and rational reconstruction) of Chomsky's position on this is that he is not interested in implementational matters, hence he does not care whether a rule is implemented as an explicit symbolic representation or an implicit, hard-wired constraint, just so long as it generates the performance regularities (grammaticality judgments) that government/binding theory is attempting to predict and explain (and there's no simpler way to do it). This, at any rate, was the reason he gave me for declining to comment on Ed Stabler's BBS target article about explicit versus implicit "representation" of grammar some years ago.

This is not the same as the "reality" issue. Nor is it the same as the explicitness/systematicity criterion for symbol systems. Chomskian rules are special cases of inferred entities; but they are noncommittal as to explicit versus implicit implementation. With symbol systems it is in order to ensure the systematicity of the semantic interpretation that representations must be explicit. Chomskian grammar (until recently, at any rate) was supposed to be autonomously syntactic, hence semantic interpretations did not enter into it. Like a thermostat (see below), a grammar allows and disallows certain FORMS of utterance; the most natural way to implement this would seem to be as a set of explicit rules operating on the shapes of the symbol strings, but there may be nonsymbolic ways to implement the rules too. Empiricism requires only that there be no alternative ways that are describable by fewer rules, or simpler ones.

Now one can say that "Strong AI" (when it claims to be doing mind modeling) is inferring the existence of theoretical entities too, namely, implemented symbol systems. That's fine, but then we need to be more specific about what it is that AI is inferring the existence of: Symbol systems are formal systems that rulefully manipulate symbol tokens in a way that is amenable to a systematic semantic interpretation. If this is the right definition of what a symbol system is, then to infer that an organism is doing something because it has a symbol system in its head is to infer that it has explicit (decomposable -- unless primitive -- and recombinable) symbol strings in its head, otherwise all the combinatary features of the semantic interpretations of the symbol combinations cannot be systematically sustained. And this is all PRIOR to any empirical evidence that the symbol systems could really generate the organism's behavior.

Substitute "sufficiently decomposable (unless primitive) and recombinable to support the systematic semantic interpretation of all the symbols, symbol strings, and symbol manipulations in the system" for "explicit" and you have the special sense of "explicit" that is at issue here. Whether a hypothesized symbol system could be edged out in its performance capacity by a rival NONsymbolic system (or another rival symbolic one, for that matter) is a different issue. And, unfortunately, looking at "innards" through neuroscience so far only seems to reveal bits and pieces of implementational detail, not a coherent view of the functional nature of the system and how it generates organisms' behavioral "competence."

> two types of thermostats... a bimetallic strip attached to a
> switch [vs.] a temperature sensor connected to a microprocessor...
> connected to a switch: ...based on the "Chomskian criterion" we would
> be justified in concluding that both of these devices explicitly follow
> the rule "if temperature is below x degrees, then turn the switch on".

No. First, as I suggested, Chomsky would not care about the difference. And second, on the basis of the systematicity criterion, NEITHER device explicitly represents that rule (unless it can do a lot more with "temperature," "below," "x," "degrees," "turn on," "switch," etc.). Rules, like semantic interpretations, must not be isolated entities but systematic ones. In isolation a stone can be interpreted as following the rule "remain stationary till pushed" (or even, if we wax fanciful, "Cogito Ergo Sum"). It is such spurious overinterpretations that the explicitness/systematicity criterion is meant to block.

> attributions of explicitness [are] convenient fictions [for] predicting
> the behavior of an organism [in] our ignorance of the actual inner
> workings of the mind/brain [or they are] simpler... in predicting...
> than a "real" description of its mechanism... [A] realist would say
> [the same] of the theoretical entities of *any* empirical science...

There's the conflation again: Explicit rules and representations are things that a system has to have in order to be a symbol system (be it merely formal, or physically implemented). Mind/brains are a special kind of system, a kind that may or may not have parts that are symbol systems (and if they do have such parts, they will have to be grounded). And the reality of theoretical entities is still another matter, independent of the question of explicitness. Explicitness is not the same as realism; empirical regularities are not the same as systematicity; semantic interpretation is not the same as theoretical inference. -- And not all things are as they seem in the hermeneutic hall of mirrors, in which we may project our semantic interpretations onto things that are scarcely capable of supporting them.

Stevan Harnad

From: Dean Z. Douthat Subject: Homely examples

It seems to me that the Chinese Room gedanken experiment itself, while cogent, powerful and devastating to symbolic AI, is rather fanciful. This leads to the unfortunate side effect of ever more "ungrounded" sci-fi fantasies. Perhaps some more homely examples would help to "ground" the argument itself.

Andy Williams has recently released an album in which all cuts are of the same song but in different languages. He learned to phonetically mimic these lyrics in only a few weeks. As he readily admits, he has no understanding whatsoever of these lyrics in any of the languages except English.

If this example seems too trivial, here's another. Opera singers frequently learn entire roles including leading roles in exactly the same way, again with no understanding of Italian, German or whatever. They are able to pass as characters in the setting, often with thunderous approval from native speakers.

Granted, these examples are not as rich and flexible as the Chinese room but they have ACTUALLY happened. Fanciful explanations seem more out of place here. Moreover, there is absolutely nothing else but the singer, no "system" to save the day -- no room, no books. Nor do singers exhibit psychological pathologies even temporarily during performances.

D. Douthat

From: Michael G Dyer To: Paul Kube (kube@cs.ucsd.edu) Subject: explicit rules

To have a physical symbol system, the implementation of the rules cannot be symbolic (o.w. as [Kube] pointed out, you get an infinite regression). If one looks at a computer, it’s clear that the circuitry does NOT follow rules, but simply behaves as though it is following rules, i.e. in a rule-describable way.

A computer gains its universality, NOT because the same hardware can act like different hardware by "following" the rules (software programs) placed in it. A computer gains its universality because laying in the software actually CHANGES THE HARDWARE, causing the computer to become a DIFFERENT machine.

That is, two identical, say, Mac-IIs, when loaded with different software, are ACTUALLY TWO DIFFERENT HARDWARE MACHINES (because "loading" the software actually makes physical changes to the memory registers, thus causing each machine to have DIFFERING HARDWARE). Sure, the Arith/logic part of the circuitry is (usually) fixed on von neumann machines, but the memory is hardware also and is part of the total circuitry controling the behavior of the machine. Computers are universal because loading software on to them makes them into different machines: A and B

Now, when we TALK ABOUT the behavior of A and B, we can EXPLAIN the differences in behavior by saying that these differences result from the different software (programs, rules, etc) in the machine, but that is NOT the real reason for the difference in behavior. The real reason is that A and B are now distinct physical machines, and thus will behave differently when turned on, due to their distinct circuitry, inexorably behaving according to the laws of electricity, physics, circuitry configuration of each.

Thus, there is no real paradox here. Computers are universal, not because of some magic "rules" that they "follow" (and therefore get into a discussion of the difference between "following a rule" and "acting as if in accordance with a rule") but because they have plasticity at the hardware level, thus becoming different forms of hardware, each of which acts as if it is following rules (the one set up by the structure of the software, which is a systematic description of how to alter the hardware before letting the machine execute), but really the machine is JUST EXECUTING.

So we can talk AS IF the machine is "reading a rule, executing, reading the next" but at the bottom level, it’s just electrons moving down wires... That is not to say that we should give up our languages (i.e. software) for describing what machines do.

So what’s a symbol system and why is it different from a feedforward PDP net or a hopfield net that is just settling? It has to do with the ability to modify the hardware of the system with input that specifies systematic ways of dealing with other inputs (these things, of course, are normally called "rules", and to handle a potentially infinite number of inputs, the rules normally have variables, and to build a potnetially infinite number of new structures, they have constituents and recursion).

Unless you can dynamically grow the memory (the tape on the turing machine; the heap in Lisp, etc.) then it's all finite. In humans, you can store memories in the environment. In computers, you can also do this (storing and retrieving from disks).

So, to Harnad's list I would add something about plastic hardware and memory management (as new memories are added) and describe the explicitness of rules in terms of how input the system (as rules) causes systematic alterations to the hardware.

"Systematic" of course, can allow a set of experiences to incrementally create something that would behave as a rule. Connectionist systems that can perform in a rule-governed way would then be symbolic also. Humans are a connectionist system that can acquire rules both from language descriptions of them, and by inducing them from experience.

Michael Dyer

From: Stevan Harnad Re: Mike Dyer on Explicit Rules

Too much about humans and computer hardware in all this. We're just trying to define a symbol system, irrespective of how it's implemented, or whether or not a human being might be one.

It seems to me the reason one wants to insist that rules must be explicit is to make sure you really have the combinatory systematicity that your interpretations are projecting onto them. Nothing to do with memory, modifiability, etc.

Stevan Harnad

From: Drew McDermott To: harnad@clarity

TO THE "TRUE SYSTEMS-REPLIER": ON THE RIGHTS AND WRONGS OF SYMBOL SYSTEMS From: Stevan Harnad I discern no argument whatsoever above, just another reflexive repetition of the original hypothesis ("if a system passes the TT, it MUST understand") in the face of the evidence that refutes it (the only system in sight, Searle, does NOT understand), with some suggestions that to think otherwise is to commit metaphysics...

I have said elsewhere ("The Red Herring Turing Test") that attributing belief in the TT to cognitive scientists is unfair. (Even more so since many eagerly succumb to being duped!) Strong AI can be phrased so it does not depend on the Turing Test. In fact, it normally is so phrased: A system that does such-and-such computations would be thinking. Of course, we can't fill in the blanks ("such-and-such" and "thinking"), but the situation is analogous to biology around 1900. In those days, "strong nonvitalists" believed that "a system that does such-and-such chemical reactions would be living." That position seemed preposterous to the Searles and Harnads of that era. If you ask the strong-AI proponent, What *evidence* is there that a system doing "such-and-such" would be thinking, the answer is, of course, Almost none! Strong AI is not based on a certain Gedanken experiment ("It's oh-so-easy to picture an intelligent robot passing the Turing test"). It's based, more or less, on faith and a vivid imagination, useful in extrapolating from today's computers.

Anyway, if we eliminate references to the TT, Searle's argument boils down to this:

1. Suppose Strong AI were true 2. Then a system that was already thinking could, by doing such-and-such computations, cause to exist, in the same hardware, two different thinking beings. 3. But that's absurd.

Obviously, those who find (2) absurd are more or less those who already found (1) absurd; and those who find (1) plausible find (2) merely an amusing, slightly unexpected, consequence of (1).

It's just an error to focus on the lack of evidence for (2). The assumption of (1) provides all the evidence we need.

I don't mean to imply that Strong AI gains credibility from the triumph of previous implausible positions, like nonvitalism in biology. Strong AI is so implausible that it's strange someone should feel the need to argue against it. I guess it's because people like me believe it in spite of its improbability.

I also don't mean to imply that belief in strong AI is irrational. It's a working hypothesis, the sort of thing that can't possibly be confirmed unless a group of people dedicate themselves to assuming it's true in the absence of evidence. What's amazing is the the number of people who seem to think it's already been confirmed. I can see why you and Searle would be annoyed by that phenomenon.
-- Drew

> Strong AI can be phrased so it does not depend on... "The Red Herring
> Turing Test": A system that does such-and-such computations would be
> thinking... "strong nonvitalists" believed that "a system that does
> such-and-such chemical reactions would be living." That position seemed
> preposterous to the Searles and Harnads of that era... I don't mean to
> imply that Strong AI gains credibility from the triumph of previous
> implausible positions, like nonvitalism in biology...

I ought to be more magnanimous and pass this one up; after all, Drew McDermott's is definitely an enlightened position. But rationality forces me to point out that:

(1) The comparison with vitalism is a red herring, and even has it backwards (I'm too much of a NONvitalist to believe in Strong AI): No one is denying that mental processes are physical processes, just that symbol manipulation is the RIGHT KIND of physical process.

(2) There's nothing wrong with the (Total) Turing Test (TTT): The problem is only with pure symbol manipulation (and the ambiguity of the symbolic Turing Test, TT). But without the TT, Strong AI is not even a "working hypothesis," it's just a(n arbitrary) definition or stipulation: For without the TT, what would count for evidence that "A system that does such and such computations would" NOT "be thinking"?

Stevan Harnad

From: miken@ai.mit.edu (Michael N. Nitabach)

SH> I have no idea what "modes of givenness" or their analysis might amount SH> to. But I have tried to show that pure symbol systems are ungrounded. SH> And that sounds like "a fundamental flaw in all symbolist theories of SH> mind" to me, at least those (like Strong AI and the Systems Reply SH> you've heard so much about in this discussion) that insist that the SH> Chinese Room DOES understand.

ME> the Chinese Room is arguing against a view of cognition that is not ME> really held by any serious cognitive scientist, even if they are ME> rightly placed in the "symbolist" camp.

SH> I guess you either haven't been following this discussion or you don't SH> consider any of my opponents serious; but let me assure you that SH> plenty of serious people do hold the view in question (and that it SH> won't be the first time that serious people have been wrong...).

As I stated early in my posting, "modes of givenness" is simply Husserl's term for the various species of functional relation a cognizant entity can bear to a symbolic representation: belief, desire, hope, anticipation, etc. Although it may have been a mistake to bring in a historical context, I think my motivation was appropriate--to suggest that much of the present discussion has been anticipated by Husserl's theory of cognition. Husserl saw that a pure symbol system cannot instantiate understanding. Furthermore, I think Husserl's resolution--to recognize that pure representation cannot explain cognition, and that an analysis of the functional relations that cognizers bear to symbolic representations is necessary-- is eminently palatable to both of us. Your theory of symbol grounding includes attempts to explain some of these functional relations. A major difference between your approach and his, however, is that he believed that those two elements (representation and functional relation) were sufficient to explain cognition, whereas you require a third element (an appropriate causal relation between events in the real world and representations). At least two other recent works (these are the only ones I have read, although I would be interested in any references you have to others) also recognize the importance of such a causal grounding for symbolic representations: Fodor's Psychosemantics, and Dan Lloyd's Simple Minds. (I am interested in how you see the relation between your notions of symbol grounding, and the ones presented in these works.)

So, I agree that "pure symbol systems" are not sufficient to provide "understanding", in any sufficiently rich sense of the term. I suppose that is so obvious to me that I lost sight of the fact that serious cog- nitive scientists could believe otherwise. My main point in that posting was to suggest that some arguments for and against the idea that a "pure symbol system" could instantiate understanding (1) have been addressed long ago, and (2) have been resolved in a way which has profound methodo- logical implications for the cognitive sciences, including AI. These implications are twofold. First, we must devote effort to the analysis of the functional relations that cognizers bear to representations. What *are* the functional relations that we label "belief", "desire", etc? Second, as you and others have emphasized, we must seek to understand the causal relations between representations and the events in the real world that they represent. Why is it that a representation represents the event that it does, and not some other? Seeking the answers to these questions-- both through study of biological cognizing systems and through attempts to create artificial cognizing systems--is much more interesting than debating whether a system which lacks both appropriate functional relations and causal grounding can "understand". Finally, I am suggesting that the works of Edmund Husserl (particularly Cartesian Meditations and Ideas) provide a *very* useful context for any present day cognitive scientist who seeks to understand Representational Theories of Mind.

Mike Nitabach

From: Stevan Harnad

Just about everyone who is not a solipsist (not just Jerry Fodor) has payed some lip service to "causal" grounding, even proponents of Strong AI (who have said all along that at some point the top-down approach requires the symbol-cruncher to be hooked up to the world via transducer/effectors "in the right way").

Let's be clear on what I, at least, mean by "grounding," and then everyone can line up under his favorite banner: Primitive symbols are labels of sensory categories. If "cat," "mat," and "on" happen to be among these, then a grounded system is one that can pick out cats, mats, and cats-being-on-mats from the projections of those objects and states of affairs on their sensors. The connection between those otherwise meaningless labels and the perceptual wherewithal to pick out their referents in the world is the kind of causal connection *I* mean when I speak of causal grounding (and in "Categorical Perception: The Groundwork of Cognition" I have a chapter in which I describe one possible model for accomplishing this).

By contrast, Fodor's "causal connection" is so vague that I haven't the faintest idea what it means, although I do know he rejects my proposal because of the "vanishing intersections" argument (which I in turn reject with a counterargument). Husserl's call for an "analysis of the functional relations that cognizers bear to representations" strikes me as equally vague (and wrong, in that the requisite relation is between organisms and objects -- the TTT -- not between cognizers and representations). I don't know Dan Lloyd's grounding scheme, or even that he has one.

The idea that sensory intersections vanish was based on introspective and a priori considerations. I think the substantive contributions to the grounding problem will come, not from past or present philosophers, but from those who face up to the constraints of TTT modeling.

Stevan Harnad

From: chris wood

On the one hand, the more I read of the symbol grounding interchange, the more I'm convinced that your position is the right one, both with respect to Searle and with respect to the "true systems repliers". I think they (the latter) have been seduced by the (real and imagined) generality of Universal Turing Machines into believing (assuming) that ANYTHING can simluated/duplicated to ANY level of specificity by a suitably-designed symbol-cruncher. Admittedly, I have trouble avoiding that particular seduction (perhaps it's a peculiar cognitive-science example of the Kahneman-Tversky "cognitive illusions" in the sense that it's hard to avoid, even when you know it's wrong; [one of their classic examples is the case in which the joint probability of events A and B is perceived, even by knowledgeable subjects, as more likely that either A or B alone]). I find the examples of airplane-simulations-not-flying and meteorlogical-simulations-not-raining-on-the-computer-room-floor reasonable antidotes to such "illusions" but the latter remain strong nonetheless.

I also think that they "true systems repliers" may have a (relatively) graceful retreat from your/Searle's onslaught by means of analysis of brains deprived of their inputs/outputs. I know this tact has been taken in earlier elements of the interchange, but I don't think it's been adequately developed. One very interesting question it raises (to me at least) is whether the necessity for symbol-grounding-in-the- world (with which I agree, as I stated above) applies only during "development" (i.e., so that, once an appropriately grounded system had "developed" during ontogeny it could be "cut loose" from the world as it were by depriving it of inputs/outputs) or whether it applies throughout the "life" of such a system (i.e., much every symbol, in order to be useful, be "grounded" via the appropriate inputs/outputs at every opportunity for its use?).

The preceding is certainly NOT for posting, but I would appreciate knowing whether you think it's a useful avenue for me to develop more thoroughly. I'd appreciate your advice in two senses: (1) Because I missed some of the interchange I'm not in fact certain how much the grounding-for-all-time issue has been discussed in earlier interchanges; and (2) You can tell me why I'd be wasting my time to pursue such an avenue of reply and save us both time and the commentary of more less-than-constructive interchange!

From: chris wood

> "true systems repliers" I think have been seduced by the (real and
> imagined) generality of Universal Turing Machines into believing
> (assuming) that ANYTHING can be simluated/duplicated to ANY level of
> specificity by a suitably-designed symbol-cruncher... I find the
> examples of airplane-simulations-not-flying and meteorlogical-
> simulations-not-raining-on-the-computer-room-floor reasonable antidotes
> to such "illusions" but the latter remain strong nonetheless.

You've slightly misstated it, and thereby almost gotten it wrong: You conflate two things. I accept the Church-Turing thesis: I take it to be TRUE that any physical process (and any abstract "effective procedure") can be simulated as closely as you like by symbol-crunching. What I deny is that simulation is the same as implementation. A symbolic simulation is merely INTERPRETABLE as being like the thing or process being simulated. It is EQUIVALENT to it under some mapping or translation ("Turing Equivalence"), but it is not IDENTICAL to it. This is where the simulated airplane and simulated thunderstorm are relevant. For there are some kinds of processes that can be both simulated AND implemented as symbol crunching, and others that can only be simulated but not implemented as symbol crunching. "Functionalism" -- or rather the variety I've called "symbolic functionalism" -- is simply blind to this difference, whereas "robotic functionalism" (and the TTT) which admits nonsymbolic functions, is not.

So the bottom line is that no matter how closely symbol manipulation can approximate an essentially nonsymbolic process like transduction, it NEVER duplicates it -- can't even come close!

Besides, the real seductive power doesn't come from the power of digital and symbolic approximations but from what I've called the "hermenuetic hall of mirrors" created by projecting our semantic interpretations onto them.

[The simulation/implementation issue, by the way, is fully discussed in "Minds, Machines and Searle."]

> "true systems repliers" may have a (relatively) graceful retreat from
> your/Searle's onslaught by means of analysis of brains deprived of
> their inputs/outputs... question whether the necessity for
> symbol-grounding-in-the-world... applies only during "development" [or]
> an appropriately grounded system "developed" during ontogeny could be
> "cut loose" from the world (i.e., must every symbol, in order to be
> useful, be "grounded" via the appropriate inputs/outputs at every
> opportunity for its use?).

This "brain-in-a-vat" issue is also fully treated in MMS. In brief, you know as well as I do that when you cut off the sensory surfaces from a brain (or simply allow it no inputs/outputs) you're still not dealing with anything like a pure symbol cruncher. Most of the brain is sensory/motor, not just its sensory surfaces. And all I've suggested is that you surely need all that nonsymbolic equipment to pass the TTT, and I'll bet you need it to pass the TT as well. How much of it could "in principle" be replaced by symbol crunchers? Not much, I bet. I'm sure a lot of it is analog in essential ways.

So that's no way out for the systems-replier. By the way, development is a red herring. Real-time history doesn't matter any more than current sensory input does. What matters is the nonsymbolic wherewithal to be ABLE to pass the TTT. Someone with a trained adult brain could in principle spring fully developed from the head of Zeus (or an AI workbench). A virtual history is as good as a real one. What matters is having the sensory-motor equipment to pass the TTT if someone ever did give you back your sensory/motor surfaces or your input/output channels.

Grounding is not magic. To really understand "The cat is on the mat," you just have to have whatever it takes to discriminate and identify cats, mats, and cats-being-on-mats if ever you should encounter them (TTT); given that (and a lot more of the same), you can certainly discourse coherently about them (TT). Neither your history nor your input/outputs need have anything to do with it (in principle).

Cheers, Stevan

THE ENTRY-POINT PROBLEM, OR WHY/HOW I AM NOT A POSITIVIST

From: kirlik%chmsr@gatech.edu (Alex Kirlik)

I still am of the opinion that your approach to the symbol grounding problem will unfortunately suffer from the same problems that plagued the positivists attempt to "ground" the meaning of theoretical terms (in a scientific theory) in the "observation terms." I do recognize that your iconic and categorical primitives are manipulated by more than their "shape," (i.e., they are truly grounded in that they are in some way "isomorphic" to pieces of the world), and that this property does indeed distinguish them from "purely symbolic" tokens. But I do not think that this distinction can deal with the problems pointed out by those critical of the positivist enterprise.

I must briefly outline the problem faced by the logical positivists. According to them, a scientific theory was to be analyzed in terms of a deductive system expressed in a formal axiomatic system or calculus. Their problem was to identify how the theoretical terms in such a symbolic representation of a theory can get their meanings. Their solution to this problem was that theoretical terms get their meanings due to their relationships with terms indicating purely observational facts and relations. The links between the symbolic elements and the observational terms were called bridge laws, correspondence rules, or dictionary statements. These bridge laws allowed for the interpretation of the theoretical terms in terms of world features directly available to perception.

Now it seems to me that, if we substitute your "symbol system" for the positivists' "scientific theory," you are both trying to solve the same problem; namely, how do the symbolic elements in these formal structures get their meanings. And you both come up with the same solution: these symbols get their meanings by being appropriately connected with lower-level "elements" (for you: iconic and categorical representations - for the positivists: observation terms) that are in some hopefully priveleged and direct contact with raw environmental features.

But the positivists ran into some serious problems with their approach, and these problems resulted in the collapse of their entire philosophy of science. There were two general problems.

The first was the impossibility of defining what counted as an observational term. I won't dwell on this although it caused them some tremendous difficulties. For your enterprise, this problem would amount to answering the question: what counts as a perceptual primitive? I would expect that you, like the positivists, would simply suggest that this can be eventually worked out, but in principle it can be done. I am not so sure that I agree (as this is one of the stickiest problems in psychology), but I will not make much of this problem here.

A perhaps more serious problem, first discussed by Putnam in his famous paper "What Theories are Not" ... contends that the positivist view that meanings for theoretical terms is drawn up, "by osmosis" as it were, from observational terms, presents a drastically innacurate view of scientific theories. Rather, he suggests that justification in science proceeds in "any direction that may be handy." He rejects the view that a scientific theory disconnected from observations is just a meaningless logical structure in need of interpretation. Perhaps the problem was best expressed by Feyerabend:

According to the view I am advocating, the meaning of observation sentences is determined by the theories to which they are connected. Theories are meaningful independent of observations; observational statements are not meaningful unless they have been connected with theories ... It is therefore the *observation sentence* that is in need of interpretation, and *not* the theory.

Putnam and Feyerabend argue that the positivists, in essence, have the whole problem turned upside-down.

It is not my point here to review and re-make the anti-positivist argument. I simply want to suggest that history has not been at all favorable to their enterprise. Due to the (I believe, very close) similarity between their approach and your own, I suggest that you will eventually encounter similar objections to your approach to the symbol grounding problem. Yes, one is a problem in Philosophy of Science, and the other of Psychology, but it's the same old Empiricist problem.

I do not expect you to have time to counter every objection to your theory that you receive. Perhaps, though, you might simply tell me whether or not you recognize the validity of the problem I have raised?

Alex Kirlik

From: Stevan Harnad

First, let me commend you on the scholarly way you presented your points. Although I had to delete the references and some other nonsubstnative passages in reposting it to the Symbol Grounding Discussion group to save space, I appreciate the care you have taken.

You have anticipated some of the responses I would make to these objections, but I don't think you have given them enough weight. There is a vast difference between the project of explicating scientific theory and of explaining how a robot (like ourselves) might work. We must be robots before we can be scientists, and we must explain our robotic capacities before we can explain our scientific capacities.

Besides the symbol grounding problem, which, I think, captures the distinction between these two explanatory projects quite perspicuously -- grounding the symbols inside a robot requires connecting them somehow to the objects and states of affairs they refer to, whereas grounding the symbols in a scientific theory already presupposes that our robotic capacities are up and around -- but there is a related problem I have pointed out now and then that I've dubbed the "entry point problem":

One cannot carve our cognitive capacities into arbitrary chunks or modules and consider them in isolation, as if they were not embedded in a larger system, with a history as well as a lot of other interdependent capacities, many of which the chunk in question may draw upon. Focusing on the theoretical vocabulary of science is an example of such an entry point problem, where one simply steps into the cognitive system at some arbitrary point far upstream from where the real action (and the grounding) is. There are examples of the entry point problem in psychology itself, for example, the Roschian work on categorization, where our ability to identify a robin as a bird was looked at very closely, but our ability to understand the words "robin" and "bird," and our ability to pick out their referents was taken for granted.

In that unfortunate line of research certain "drastic" conclusions were also drawn, for example, that the "classical view" of categorization -- the one according to which objects are categorized on the basis of common features -- was "seriously wrong," and that category membership was really just a matter of degree.

Well, I'm inclined to mistrust received views according to which certain approaches have been decisively discredited unless the fatal problem with the approach in question is a clearly demonstrable logical or empirical (or probabilisic) one. (I've seen too many pendula swing back and forth, and too many bandwagons come and go.)

There has been no such clear demonstration of a fatal problem with certain forms of empiricism. On the other hand, I'm not a positivist or a verificationist. In fact, I'm not really proposing a theory of "meaning" at all (although I do use the word "meaningless" a good deal in describing ungrounded symbol systems). I am really only proposing a theory of GROUNDING: A way that semantically interpretable symbols inside a robot could be connected (functionally and causally) to the objects and states of affairs that they refer to in the world the robot lives in. It's no use telling me about the "coherence" of the symbols inside the robot: I want to know how they succeed in picking out the objects they refer to. And positivism or no positivism, this has to have something to do with the sensorimotor system.

Let me say the rest in the context of the specific points you raise:

> According to [the logical positivists] a scientific theory was...
> a deductive system... in a formal axiomatic system...
> how [do] the theoretical terms... get their meanings... Solution:
> [from] their relationships with terms indicating purely observational
> facts and relations. The links... were called bridge laws,
> correspondence rules, or dictionary statements [which] allowed for the
> interpretation of the theoretical terms in terms of world features
> directly available to perception.

For me, the logical positivists' enterprise was from the start pilloried on the entry point problem, which, as I've suggested, is one of the horns of the symbol grounding problem: Not only are the theoretical terms of science too far upstream to start with, but, as my "Chinese/Chinese Dictionary-Go-Round" suggests, symbolic rules, dictionary definitions and interpretations won't help, for they are every bit as ungrounded (and hence derivative) as the theoretical terms themselves. This is even true of the positivists' most unproblematic kind of "observation statement," such as "Snow white now," for it simply takes for granted the sensory and cogntive wherewithal to pick out "snow," "white," "now," and "snow's-being-white-now," whereas explaining THAT is what I take the real problem to be!

> you both come up with the same solution: these symbols get their
> meanings by being appropriately connected with lower-level "elements"
> (for you: iconic and categorical representations - for the
> positivists: observation terms) that are in some hopefully priveleged
> and direct contact with raw environmental features.

But don't you see that there's a world of difference between "observation terms" -- which are just more symbols -- and the nonsymbolic structures and processes I'm talking about? In the Chinese/Chinese Dictionary, your simple, concrete sensory terms are every bit as ungrounded as goodness, truth, beauty, quarks or superstrings. And it's simply fantasy to think that their successful link to whatever it is that they refer to can just be taken for granted -- a simple transducer hook-up. In fact, I claim that, if the motivation for the positivist program came from some aspect of the symbol grounding problem, the motivation for the critiques of positivism did too! Not only are sensory symbols every bit as ungrounded as abstract ones, but they are almost equally theoretical. I can go on spinning out the definitions and interdependencies of my concrete sensory vocabulary as endlessly as I can those of my abstract and theoretical vocabulary. It's just the symbol/symbol dictionary-go-round either way.

But for a real robot, the buck has to stop somewhere, and I'm suggesting it stops with the functional wherewithal to pick out the things the symbols refer to. (This is accomplished, according to my theory, by the iconic and categorical representations.) I don't care about the "theory-ladenness" of "The snow is white" as long as the robot can pick out the objects and states of affairs referred to. Moreover, in insisting on the "Total Turing Test" (TTT) as the criterion, I'm avoiding other manifestations of the entry point problem, in the form of trivial toy robots that only exhibit an arbitrary chunk of capacity.

Now if the meaning theorist objects that a TTT-scale robot may still not have captured "meaning," I'm content to say that groundedness was all I ever aspired to, and it's sure got that!

> Serious problems: [1] impossibility of defining what counted as an
> observational term [=] ...what counts as a perceptual primitive?

I happen to have put my money on categorical perception as the "perceptual primitive," for reasons I decsribe and marshall evidence for in "Categorical Perception: The Groundwork of Cognition" (Cambridge University Press 1987, S. Harnad, ed.), but THAT there are elementary perceptual categories we can identify and name is, I think, an empirical fact we all know at first hand. My mission is to explain how we do it, and there's no definitional problem whatsoever: Perceptual primitives are just the smallest categories we are able to identify and manipulate reliably. No one is talking about DEFINING them, just picking them out.

Now, once a "basis set" of labelled categories has been picked out, it is possible to "define" new, undefined labels by combining and recombining the grounded primitives. The procedure can be more complicated (there can be bootstrapping and revision of primitives, and coherent interrelationships must be preserved, at least to an approximation), but those are details (discussed in "The Symbol Grounding Problem" and the book I mentioned), whereas the grounding proposal itself is quite clear and simple.

I don't have to specify what the actual primitives are (once grounded, they're probably in some degree of actual or potential flux under the influence of new data, both sensory and symbolic); I can't even say how many is enough. But there certainly seems to be no problem with the inference that they exist, as long as you don't demand an autonomous and immutable definition of them. I have no temptation to define them, because definitions already presuppose a grounded symbol system, whereas I'm just trying to get one of the ground in the first place.

> [That meaning is] drawn up, "by osmosis" [is] a drastically innacurate
> view of scientific theories. Rather justification in science proceeds
> in "any direction that may be handy." [Putnam] rejects the view that a
> scientific theory disconnected from observations is just a meaningless
> logical structure in need of interpretation.

Once a symbol system is grounded, revision can proceed in all directions, including top-down. (Primitives are just a provisional basis set, not a lapidary "foundation" -- that is, if you're only thinking of robotics rather than philosophy of science or ontology.) But the grounding itself can only come from one direction: bottom-up. Otherwise, symbol systems, be they ever so coherent and interpretable, are simply hanging from a skyhook, with their interpretations completely parasitic on our projections (the "hermeneutic hall of mirrors"). And, as I've suggested, scientific theory is much too far downstream to serve as a representative case.

> Feyerabend: "meaning of observation sentences is determined by the theories
> to which they are connected. Theories are meaningful independent of
> observations; observational statements are not meaningful unless they
> have been connected with theories ... It is therefore the *observation
> sentence* that is in need of interpretation, and *not* the theory"...
> the positivists, in essence, have the whole problem turned upside-down.

By now I think I've already replied to this. Both theories and observation statements are ungrounded strings of symbols whose meaningfulness is projected onto them by our minds. The symbols in our heads must be grounded by the nonsymbolic structures and processes that allow us to pick out (enough of) the things the symbols stand for. Symbol-coherence theorists are trapped in the hermeneutic circle, but robots are not.

> Due to the (I believe, very close) similarity between their approach
> and your own, I suggest that you will eventually encounter similar
> objections to your approach to the symbol grounding problem. Yes, one
> is a problem in Philosophy of Science, and the other of Psychology, but
> it's the same old Empiricist problem.

I've answered this too, I trust. The right entry-point to this problem is a robotic theory, not philosophy of science.

Stevan Harnad

EXCHANGE WITH NITABACH AND MCDERMOTT

From: miken@ai.mit.edu (Michael N. Nitabach) Subject: Defining Symbol Systems

Thank you for clarifying the technical meaning of the explicit- implicit distinction. Clearly, in my post I was actually discussing a different issue--the reality issue. So, there are actually three theoretical "axes" here. Explicit-implicit with regard to systematicity, explicit-implicit with regard to rule implementation, and real-unreal with regard to theoretical entities. I think that the systematicity issue is the least interesting of the three, at least from an empirical point of view. E.G. Dan Dennett is at great pains to deny the reality of intentionality as a characteristic of the mental; he regards it as "way of talking" about complex systems. The explicitness of implementation of rule following is an important issue today, especially given much controversy about connectionist nets as cognitive models. I think the nets will be seen as poor models of symbolic processes (like language), but will be key elements of theories of the non-symbolic transduction-grounding processes.

With regard to my other posting (on causal grounding), I agree with you that the resolution of the "vanishing intersections" issue can only come through empirical research into the possibility of creating devices which achieve grounding through the mechanisms you propose. I have to disagree, though, with your claim that "substantive contributions to the grounding problem will come, not from past or present philosophers, but from those who face up to the constraints of TTT modeling." As I said, I do agree that empirical work is necessary to resolve the "vanishing intersections" issue. However, there is also a lot of "philosophical" work to be done as well-- e.g. clarifying terms and exposing logical flaws in proposed explanations concurrently with the empirical work. This is just a matter of pure speculation on my part, though. However, I think you cannot deny the strong connection between your theory of causal grounding and the previous attempts at the formulation of a procedural semantics, by philosophers. The essence of procedural semantics is the notion that for a symbol to be grounded is to be in possession of a procedure for picking out its referent based on sensory input. So, you say: "[A] grounded system is one that can pick out cats, mats, and cats-being-on-mats from the projection of those objects and states of affairs on their sensors. The connection between those otherwise meaningless labels and the perceptual wherewithal to pick out their referents in the world is the kind of causal connection *I* mean when I speak of causal grounding." This is exactly the claim of procedural semantics. I have read all the papers you have sent me about your theory of symbol grounding, and I still don't see how it goes any farther than what has already been proposed as procedural semantics. Note, that I am not making a claim here about the tenability of procedural semantics, and hence your theory of grounding. My point is just that your theory does not exist in a conceptual vacuum, but actually follows very closely a tradition that was developed in large part by philosophers.

--Mike

From: Stevan Harnad

If procedural semanticists say that symbols are grounded by whatever structures and processes pick out their referents from sensory input then I agree with them. I have gone on to suggest what those structures and processes might be, however, namely, iconic representations that allow us to discriminate and categorical representations that allow us to identify objects, properties and states of affairs by finding the invariant features in their sensory projections (perhaps through connectionist learning); the names of the elementary sensory categories then become the elementary symbols in a grounded symbol system that generates the rest by the composition of grounded categories. That's not philosophy, it's testable hypotheses (and they are currently being tested by several different groups, through both behavioral experiments and computer modeling). (I agree that good philosophers can help, by clarifying concepts and exposing logical flaws.)

Stevan Harnad

From: Drew McDermott (1) SH: The comparison with vitalism is a red herring, and even has it backwards (I'm too much of a NONvitalist to believe in Strong AI): No one is denying that mental processes are physical processes, just that symbol manipulation is the RIGHT KIND of physical process. DM: I didn't mean that anti-computationalists are literally vitalists. It was an analogy to a previous scientific controversy.

(2) SH: There's nothing wrong with the (Total) Turing Test (TTT): The problem is only with pure symbol manipulation (and the ambiguity of the symbolic Turing Test, TT). But without the TT, Strong AI is not even a "working hypothesis," it's just a(n arbitrary) definition or stipulation: For without the TT, what would count for evidence that "A system that does such and such computations would" NOT "be thinking"? DM: Comments in reply:

(a) Odd you should phrase the question as "What would count for evidence that such-and-such computational system was NOT thinking?" Surely the burden is going to be on those who claim it IS thinking for the foreseeable future. If it ever begins to seem obvious to the majority of observers that a computer is thinking, who is going to want evidence that it isn't?

(b) Methodological assumptions like strong AI do not get refuted system-by- system. We're not going to wake up in the year 2100 with the task of hanging signs on the machines that only *appear* to be thinking. Such assumptions either just become part of the general background scientific culture, or they become the property of an ever-smaller minority. Take astrology, for instance, which was originally the very paradigm of an exact science, and is now believed by no real scientists, but was never actually disconfirmed. How could it be? (Astrology is an odd case in that it has retained popularity among those whose understanding of science is arrested at the A.D. 1000 level, i.e., most people.)

(c) Suppose computationalism does get confirmed in the sense that nonvitalism -- or "chemism" -- got confirmed, then questions about exactly who's thinking and who isn't are going to be as important as questions about exactly who's alive and who isn't, which is to say, not important at all. It's an interesting high-school puzzle whether viruses are alive, but since we now know that life = organic chemistry, and we understand viruses thoroughly as organic-chemical systems, no one cares whether they are alive or not. Similarly, if thinking = computing, then there will be plenty of systems that "sort of, in a way, think," just because they exhibit some aspects of thinking, even no one would be fooled into mistaking them for persons.

Drew McDermott

SH: I think this is still just an argument by analogy to vitalism, and it doesn't work, because there's something it's LIKE to be thinking, and you either have it or you don't. It does not admit of degrees.

Think about it before replying "Yes it does!" I'm not talking about how well you think, how much you think, how fast you think, how much you know, or even what the qualitative nature of your thinking is, all of which ARE amenable to questions of degree. I'm just talking about the fact THAT thinking has a qualitative nature at all; that's an all or none matter.

The same is definitely not true of living/nonliving, which, as an objective* property, may either admit of degree or slide gradually into a gray area where the distinction breaks down. This is simply not true of the subjective state of thinking. (As for "thinking" in devices that have no subjective states: I think it's arbitrary to even use the same word for it. You may as well talk about them "feeling" too.)

-----

*Not to be confused with the subjective side of life -- "what it's like to be alive" -- which is of course really a question about mind, not life.

And of course the problem of how to show that a device DOESN'T think is critical. Otherwise, like Humpty Dumpty, we'll just be saying things think because we say they think. There is, after all, a fact of the matter about whether or not a device is thinking (an all-or-none fact, as a matter of fact).

The only problem is, you must BE the device to know it for sure. Except for special cases like the Chinese Room, one rarely has the opportunity to perform such a test...

Stevan Harnad

[Some relevant cross-postings; apologies to those who have seen them already. SH.]

Subject: Re: STRONG AND WEAK AI

Chris Malcolm asked for a definition:

Those who believe in Strong AI believe that thinking is computation (i.e., symbol manipulation). Those who believe in Weak AI believe that computation is a means of studying and testing theories of (among other things) thinking, which need not be just computation (i.e., not just symbol manipulation).

The typical error of believers in Strong AI is a misconstrual of the Church-Turing Thesis: Whereas it may be true that every physical process is "equivalent" to symbol manipulation, i.e., is simulable by symbol manipulation, it is decidedly NOT true that every physical process IS symbol manipulation. Flying, heating and transduction, for example, are not. How does one fall into this error? By becoming lost in the hermeneutic hall of mirrors created by the semantic interpretations we cast onto symbol systems. We forget the difference between what is merely INTERPRETABLE as X and what really IS X. We confuse the medium with the message.

The chimpanzee language experiments (and, to a lesser degree, "Clever Hans") fell into similar errors. Freudian interpretations of the machinations of the unconscious and astrological interpretations of what the heavans portend are more distant relatives...

Stevan Harnad

Subject: Cog Sci Fi (was: STRONG AND WEAK AI) Summary: Lost In the Hermeneutic Hall of Mirrors

This is a multiple reply, to multiple errors:

cam@aipna.ed.ac.uk (Chris Malcolm) of Dept of AI, Edinburgh University, UK, wrote:

> my game is assembly robotics... The assembly agent is designed not only
> to succeed in its tasks, but to present a suitable virtual world to the
> planner, there is an extra constraint on the task modularisation. That
> constraint is sometimes referred to as the symbol grounding problem.

As the mint that coined the term, I think I speak with a little semantic authority when I say that that's decidedly *not* the symbol grounding problem but rather a symptom of it! Virtual worlds are not real worlds, and what goes on inside a simulation is just meaningless symbol crunching. The only way to ground symbols is in the real world.

Let me add the prediction that whereas a "virtual world" may allow one to ground a toy robot in the real world, it will never lead to what -- for a psychobiologist, at any rate -- is the real goal: A robot that passes the Total Turing Test. The reason is again the symbol grounding problem: Virtual worlds cannot embody all the contingencies of the real world, they can only capture as many of them symbolically as we can anticipate. In ordinary AI this was known as the "frame" problem -- but of course that's just another manifestation of the symbol grounding problem.

mike@cs.arizona.edu (Mike Coffin) of U of Arizona CS Dept, Tucson, wrote:

> an artificial intelligence living... in a (sub-)simulation on a Cray-9
> would have no choice but to accept the simulated flight to Istanbul to
> the AI conference as ‘‘reality.’’

Here is the other side of the coin, what I have called the "hermeneutic hall of mirrors" created by projecting our interpretations onto meaningless symbols. As long as you allow yourself to interpret ungrounded symbols you'll keep coming up with "virtual reality." The only trouble is, what we're after is real reality (and that distinction is being lost in the wash). There's nobody home in a symbol cruncher, and it's not because they're on a virtual flight to Istanbul! This is what I call "Cog Sci Fi."

yamauchi@cs.rochester.edu (Brian Yamauchi) of University of Rochester Computer Science Department wrote:

> My complaint about most AI programs is not the worlds are simulated,
> but that the simulated worlds often are very unlike any type of
> perceptual reality sensed by organic creatures. It's a matter of
> semantics to argue whether this is "intelligence"...
> It seems that one interesting approach to AI would be to use the
> virtual reality systems which have recently been developed as an
> environment for artificial creatures. Then they would be living in a
> simulated world, but one that was sophisticated enough to provide a
> convincing illusion for *human* perceptions.

Illusion is indeed the right word! Simulated worlds are no more "like" a reality than books are: They are merely *interpretable* by *us* as being about a world. The illusion is purely a consequence of being trapped in the hermeneutic hall of mirrors. And that's not just "semantics," it's just syntax...

Stevan Harnad

----

mnr@daisy.learning.cs.cmu.edu (Marc Ringuette) of Carnegie-Mellon University, CS/RI wrote:

> As a practicing roboticist, it's clear that when I consider an AI
> system I should not be asking the question "Is it grounded?" but rather
> the question "How interesting is it?", where _interest_ is positively
> correlated with _realism_... I think it's up to you to give reasons
> why the AI community should care about the philosophical issue of
> _grounding_ rather than the practical issues of _interest_ and
> _realism_.

As a practicing roboticist, you can be interested in whatever you like. But if you're doing real robotics, rather than virtual robotics, your robots better be able to do whatever they do in the real world. To the extent that symbol-crunching in a virtual world can actually be translated into robotic performance in the real world, none of my objections should worry you. To the extent it cannot, they should. For my part, my interest is in a robot that can pass the Total Turing Test; the symbol grounding problem for this enterprise is empirical and methodological. It isn't and never has been philosophical.

Stevan Harnad

>From: harnad@phoenix.Princeton.EDU (Stevan Harnad) Subject: Re: Cog Sci Fi (was: STRONG AND WEAK AI) Summary: Real worlds vs. virtual worlds...

yamauchi@cs.rochester.edu (Brian Yamauchi) University of Rochester Computer Science Department

> Suppose virtual reality technology develops to the point where it is
> impossible for a human to tell an illusion from reality (this is
> already the case for still computer graphic images of some objects)...
> Now, suppose that we can develop a program which can react to these
> images in the same way that a human can... Now, if you are arguing that
> it will be impossible in *practice* to build a simulator which has the
> complexity of the real-world, in terms of interactivity and modeling of
> complex physical laws, then you may have a point.

First of all, the last point WAS my point: The problem of designing a robot that will pass the Total Turing Test (TTT) is a tiny subset of the problem of simulating the world the robot is in, not vice versa. (Another way to put it is that an analog object or state of affairs is infinitely more compact than any possible symbolic description of it. To approximate it closely enough, the description quickly becomes ludicrously large.)

Second, the point of the TTT is for the ROBOT to pass it, in the world and for us, not for its WORLD to pass it for us.

Finally, it's irrelevant what graphics we hook onto one symbol cruncher in providing inputs to another symbol cruncher. (That's like having two computers play chess against one another: it may as well all be one computer.) Symbol crunchers don't see, not even if you hook transducers onto them. It's just playing on the illusion from OUR point of view to bother having a graphics interface. So it's still just the hermeneutic circle, whether we're projecting our interpretations on symbolic text or on symbol-governed graphics.

Stevan Harnad

From: harnad (Stevan Harnad) To: hendler@cs.UMD.EDU Subject: Matters of Degree...

Jim,

Please reread the message you sent to me. You have failed to understand the very point you quoted from me: The matters of degree (in goldfish, children, us) that you replied about were EXACTLY the ones I warned Drew McDermott in advance NOT to try to cite, because, though of course I agree they exist, they're completely irrelevant to the all-or-none issue of whether or not there is anything it's LIKE (subjectively) to be doing those various degrees of thinking. Where there is, it's really thinking; where there isn't, it's not.

The varying degrees of thinking CAPACITY and CONTENT in biological organisms, including ourselves, linked by physiological and phylogenetic continuity, are themselves irrelevant to the question of the putative "thinking" of a computer, which can lay no claim to being on that physiological/phylogenetic continuum.

I repeat: The degree question is not HOW MUCH you can think, but WHETHER you can think at all. Thinking is a subjective state; if there's no subjectivity, there's no thinking. Animals probably have subjectivity as we do, because they're so like us and continuous with ourselves physically, causally, functionally, historically, ecologically.

The only "lookalike" evidence I would accept for an artifact not on this continuum is the Total Turing Test. I'd accept it for an artificial goldfish too, but we do not have anywhere enough ecological knowledge or intuition to apply it to a goldfish, so we're stuck with the bigger test that's closer to home, in which we're judging the presence/absence of a subjective state (1) that we know is there in people (modulo the "other minds" problem) and that we (2) each know intimately at first hand.

Stevan Harnad

UNCOMPLEMENTED CATEGORIES

Date: Fri, 15 Dec 89 14:02:53 PST From: Pat Hayes To: Stevan Harnad [SH] Subject: Matters of Degree...

Stevan,

You insist that being a thinker is an all-or-nothing subjective state. But on what authority do you make such a claim? Just intuitively, it seems that awareness admits of degrees: and I do not mean to refer to how much one knows, or how fast one thinks, etc., but simply to the awareness itself. One can be half-awake, for example, or can be in a state of confusion after a blow on the head. But a better example is the awareness which we attribute to animals, such as our pet dogs and cats. It seems likely that they are aware of their surroundings, and that we can even communicate with them to a limited extent, but it also seems clear that their awareness is inferior to ours. What would it be like to be a chimpanzee/cat/dog/bat/vole....? Surely, the initial assumption that a moments reflection about nature would suggest must be that awareness does indeed admit of degrees, like any other quality.

[SH: To save lines I will respond after each paragraph, rather than requoting at the end. What I insist is all-or-none (and once you understand what I mean I don't think you'll be able to disagree with me) is WHETHER there is any awareness going on at all. Not how much; not of what; just whether. Degree of wakefulness, confusion, alertness, attentiveness, energy, etc., have nothing whatsoever to do with this question. (You ARE aware in several stages of sleep, by the way; there's just no continuity or memory of the awareness unless you're awakened during that stage.) It's an exclusive-or question. As to the awareness we "attribute" to animals: That can hardly help clarify matters, since it's just an attribution; we have no idea awhether we're right about what it's like for them. But if there's anything at all it's like -- and your own way of putting it confirms that we are indeed attributing AWARENESS to animals -- then again it is there; if not, it isn't. Whether "superior" or "inferior," narror or wide, faint or intense, awareness is awareness, and absence of awareness is absence of awareness. How much the awareness is awareness of, or what it's the awareness of, or even what it's really like, is all irrelevant. If there's any awareness at all, there's awareness, if there isn't, there isn't. The only other point is that "thinking" in a device that has no awareness isn't thinking (but you don't seem to be disputing that). SH.]

You briskly dismiss the living/nonliving distinction as analogous, but this most certainly was considered a very sharp distinction until a few years ago. The very idea of 'organic' chemicals is based on an idea that the living stuff was fundamentally different in its very nature than inorganic stuff: protoplasm, it was called. And even after living material was seen to be composed of chemicals, it was thought that there was some sort of sharp distinction between living and dead organisations of matter. Viruses were the oddity, and the question of whether they were alive or not was considered a real, hard, question until about 20 years ago. Being aware is surely far more a matter of degree than being alive.

[SH: Yes, I dismiss it as mere analogy, and superficial analogy at that. It's irrelevant that people thought life was an all- or-none thing. There was no logical reason it had to be. It was a hunch (probably itself parasitic on the mind/body problem) that just seems to have turned out to be wrong. Living/nonliving was always an OBJECTIVE distinction, and the objective cards are free to fall where they may -- across a clearcut dichotomy or along a graded continuum. Not so with SUBJECTIVITY itself. There either is some (be it ever so dim) or there's none. You might say it's part of the logic of the phenomenology (subjective phenomenology, of course) itself. It's part of knowing what it's like to be aware to know that it's an all-or-none phenomenon: A phenomenon that grades in terms of its intensity and extent to be sure, but not its existence. (It's not surprising that we have problems getting a grip on this peculiarity of subjectivity. It is peculiar among our concepts because it is "uncomplemented" -- it has no negative instances, i.e., there is NOTHING it's "like" to be nonaware (not to be confused with being aware, but unaware of, say, X) -- as I have pointed out in a paper called "Uncomplemented Categories, or What is it LIke to be a Bachelor?") This invariably leads to conceptual problems, because we resort to incoherent analogies in order to supply the missing complement. SH.]

You say that computers are not on the physiologial/phylogenetic continuum as biological organisms. Thats true: but so what? The idea that computers might be, in some sense, aware, follows from a hypothesis about what constitutes awareness, viz. the appropriate possession and manipulation of suitable grounded representations of the environment. This is a hypothesis, and it isn't clear whether it is true or not. But it can't be dismissed as wrong by definition. Suppose that something passes the TT (not even the TTT), and it insists that it is aware, has feelings, has the subjective state of thinking, etc.. On whose authority would you claim that it did not *really* have the authentic subjectivity which it spoke of? My inclination would be to say that it knew more about how it felt than you or I do. Suppose for example we had such a device which was blind, deaf, etc., and could only pass the TT. Now put it through surgery and give it eyes, so it now passes (part of) the TTT. Does it suddenly have an authentic subjective state which it never had before? Suppose it can remember its blindness and can discuss its past: what is it remembering? It just doesn't seem reasonable to draw these boundaries around 'real' awareness, they will always be indeterminate.

[SH: Biological continuity is relevant and an important empirical datum. But let's set it aside for now. The hypothesis that pure symbol-manipulators (let's call a spade a spade) might be aware was definitely a tenable one, and biology was no a priori objection to it. I reject the hypothesis for the following three reasons, two negative and one positive: (1) I do not find that actual empirical work is leading in the direction of passing the TT (and less than the TT I hope neither of us would settle for: Just a string of sentences "I'm aware, yes I am, yes I am..." or what have you). (2) I acknowledge the validity of Searle's Chinese Room Argument (which shows that even if you WERE the symbol manipulating system, you wouldn't understand). (3) And I prefer the alternative hypothesis, suggested by the Symbol Grounding Problem, that pure symbol crunching is ungrounded, and hence a mere symbol cruncher cannot be aware: To be aware a system must pass the TTT, which can't be done by symbol-crunching alone; it requires a hybrid nonsymbolic/symbolic system, grounded in the world to which its symbols refer. I also think this shows why pure symbol crunching can't even pass the TT: Because successfully passing the TT would itself need to draw on the (nonsymbolic) capacity to pass the TTT. So, as I wrote in "Minds, Machines, and Searle," I do not have to commit myself to any intuitive contingency plans about what I would believe if the Sci Fi scenario in which a pure symbol cruncher passed the TT were a reality, and it implored me to believe it had a mind: (1) - (3) are empirical and logical reasons enough for believing that that Sci Fi scenario is just Sci Fi. (And, as I've said many times before, TTT-scale symbol grounding does NOT amount to just sticking trandsducers onto a pure symbol cruncher and suddenly the lights go on, voila! Nor is a pure symbol cruncher like a deafferented brain; so the deaf, dumb, blind, paralyzed analogies are all gratuitous projections, symptoms of being trapped in that ubiquitous hermeneutic hall of mirrors...) SH.]

You insist on the attribution of mental states to a computational device as being simply an error, a hermeneutical reflection. But this isn't justified. It's not a mistake, it's a claim: and you can't just dismiss it, you need to argue against it. And you need to do better than John Searle.

[SH: But I have argued against it. And although I think John Searle has done quite well, I think I've taken the ball still further, pinpointing just what it is about symbol crunching that's responsible for both its empirical performance limitations and its vulnerability to the Chinese-Room Argument (namely, the symbol grounding problem) and suggesting an alternative (a hybrid system). And whereas all lookalikes (including TTT-scale ones) might be just projections (because the other-minds problem is not solvable, and will always leave an empirical theory of mind with a level of underdetermination that is at least an order of magnitude greater than the underdetermination of theoretical physics), symbol systems are especially prone to being drawn into the hermeneutic circle: That's what the interpretation of symbols is, after all. SH.]

PS. Your definitions of strong and weak AI aren't quite fair. The terms are due to Searle, and my understanding of the distinction is that "weak" AI claims that AI programs are to be thought of as simulations of thinking: models, perhaps, like simulations of the weather; whereas "strong" AI claims that the programs are actually doing thinking, perceiving etc.. The core of the distinction is whether computational models are merely simulation or are in fact examples of cognition. Now, as I understand it, this distinction says nothing about whether the computational model is purely symbol manipulation. Strong AI allows for all sorts of computation, and also allows for - some would even insist on - there being perceptual links to the external world. Its not a question of the strong/weak distinction being one of restriction: strong AI insisting on only symbol manipulation, but weak AI allowing other stuff as well. --- Pat Hayes

[SH: I don't understand your point. I think you've described Searle's strong/weak AI distinction exactly the way I would. A mere computer program, doing nothing but computing, is just manipulating symbols. It may be simulating thinking, but it is not thinking, just as it may be simulating a storm, but not storming. What's going on is just ungrounded symbol manipulation, plus an interpretation projected onto it by us. Now what are these other forms of "computation" you have in mind, that are not just symbol manipulation? And, as we discussed before, to the extent that the success of your model depends functionally on "perceptual links," it's not just a computer program. I mean, a computer might be playing some small role in generating real storms, but so what? The storming is still the noncomputational part. -- Stevan Harnad.]

From: anwst@unix.cis.pitt.edu (Anders N. Weinstein) Organization: Univ. of Pittsburgh, Comp & Info Services

"Explicit representation of the rules" is a big red herring.

At least two major articulations of the "symbolist" position are quite clear: nothing requires a symbol system to be "rule-explicit" (governed by representations of the rules) rather than merely "rule-implicit" (operating in accordance with the rules). This point is enunciated in Fodor's _Psychosemantics_ and also in Fodor + Pylyshyn's _Cognition_ critique of connectionism. It is also true according to Haugeland's characterization of "cognitivism" [Reprinted in his _Mind Design_]

The important thing is simply that a symbol system operates by manipulating symbolic representations, as you've characterized them.

Many people seem to get needlessly hung up on this issue. My own suggestion is that the distinction is of merely heuristic value anyway -- if you're clever enough, you can probably interpret any symbol system either way -- and that nothing of philosophical interest ought to hinge on it. I believe the philosopher Robert Cummins has also published arguments to this effect, but I don't have the citations handy.

Anders Weinstein U. Pitt. Philosophy Pittsburgh, PA 15260

From: Stevan Harnad

Anders Weinstein wrote:

> "Explicit representation of the rules" is a big red herring...
> nothing requires a symbol system to be "rule-explicit" (governed by
> representations of the rules) rather than merely "rule-implicit"
> (operating in accordance with the rules). The important thing is simply
> that a symbol system operates by manipulating symbolic representations,
> as you've characterized them... if you're clever enough, you can
> probably interpret any symbol system either way

I'm not convinced. Whether a rule is explicit or implicit is not just a matter of interpretation, because only explicit rules are systematically decomposable. And to sustain a coherent interpretation, this decomposability must systematically mirror every semantic distinction that can be made in interpreting the system. Now it may be that not all features of a symbol system need to be semantically interpretable, but that's a different matter, since semantic interpretability (and its grounding) is what's at issue here. I suspect that the role of such implicit, uninterpretable "rules" would be just implementational.

Stevan Harnad

[Not Posted to the Symbol Grounding Discussion Group]

Date: Fri, 15 Dec 89 22:04:25 EST From: delta@csl.ncsu.edu (Thomas Hildebrandt) Subject: Food for thought?

It occurred to me recently, that by using the Kolmogorov decomposition, one can force a clean dichotomy between "thinking machines" and the ordinary serial computer, if one can show that the former category exists.

The Kolmogorov Decomposition Theorem states that any function of N variables can be represented as an arbitrary function of a linear combination of at most 2N+1 arbitrary functions of linear combinations of the inputs. In symbols:

Let X represent the vector (x[1], x[2], ..., x[N]). Then f(X) = g(h1(X) + h2(X) + ... + h[2N+1](X)), where h?(X) is of the form h?(X) = s?(a[1]x[1] + a[2]x[2] + ... + a[N]x[N]).

g() and the s?()s are arbitrary scaling functions.

The interesting thing about this decomposition is that it allows computation of the original N-input function using nothing but unary and binary operations (just like an every-day von Neumann style serial computer). Any N-input function can be computed using a serial computer.

Well, any deterministic function. But then, if you let a function be non-deterministic, than it isn't a function anymore, since in essence, you are letting it assume two or more values simultaneously.

So, the question is this: Are thinking machines non-deterministic? (I believe that they can be.) If not, then there is no distinction between thinking machines and ordinary computers (outside of the potential gain in speed, using a parallel implementation of the former). If so, then "intelligent behavior" depends upon nondeterminism.

It seems rather counterintuitive that the defining characteristic of intelligence is unpredictability, but there you have it.

This definition seems to match well with observation, too. Poke an earthworm (not very intelligent) any number of times, and it will shrink in almost the same way each time (very predictable). Poke a cat (which, I hope you will agree, is more intelligent than an earthworm) a number of times, and it may (a) glare at you, (b) hiss and rake the back of your hand, (c) get up and walk away, or (d) all of the above. Whether or not, and when it initiates any of these reactions will change from day to day, poke to poke, cat to cat (not very predictable).

Humans are even less predictable, to the extent that it is not considered wise to go around poking them. "Oh, excuse ME." "Ouch." "Hey! You **** **** what the **** **** do you **** **** ****!" "Biff" "Kapow!" etc.

You never know. . . . Thomas H. Hildebrandt delta@csl.ncsu.edu

----

From: Stevan Harnad

Tom:

The part about deterministic/nondeterministic systems seemed fine, but then you made a leap that was completely arbitrary: Intelligence is not to be "defined" (any more than gravity is). It’s an empirical phenomenon whose causal mechanism we wish to understand. Kolmogoroff does not help, and certainly not when it is wrapped in arbitrary interpretations of the variability and plasticity of the behavioral capacity of higher organisms. All you get out of stipulating that "unpredictability" and "nondeterminism" is what’s going on is what you put into it in the first place.

Stevan Harnad

Date: Sat, 16 Dec 89 01:19:27 EST From: delta@csl.ncsu.edu (Thomas Hildebrandt) To: Stevan Harnad

Here is the third part of the correspondence I promised you.

(3) I disagree that complete grounding of symbols is necessary to interesting, human-like applications. In fact, ungrounded symbols are necessary to human-like thought, or any kind of intelligent behavior. I assume that one interpretation of an ungrounded symbol is a variable to which no value has yet been assigned. We use such variables all the time. Where is there room in a brain for anticipation, if the value (read interpretation) of a symbol is grounded (known) at all times? How can one say, "If the total is more than the $20, I’ll have to put something back."? Where is there room in the brain for the as-yet- undetermined symbol "total"?

[SH: This is either a misunderstanding of grounding or a straw man. There’s nothing aboutmy grounding proposal that implies you can’t have unbound variable names. SH.]

At the very heart of abstract thought is the substitution of symbols for instances. In grade school, we memorize a bunch of examples like 1+2=3 4+6=10, etc. Then, in algebra, we get x+y=z. Then, it turns out that you can represent algebras and number systems symbolically. There are libraries filled with mathematical results that are WAITING to be grounded. Are they meaningless until someone finds an application? I suppose you might say so, but the person who finds the application is likely to have a different opinion on the subject. And that two people can stand and discuss mathematical theories, long before either concerns himself with an application, indicates that complete grounding is necessary to mutual understanding.

[SH: This is still just misunderstanding. Of course we can do mathematics as just rule-following; and of course we can talk about undefined symbols syntactically. But you have to be grounded to be able to do that kind of thing in the first place. Your mistake is an example of the "entry point problem" I mentioned in an earlier posting. You also don't seem to have a clear idea of the rule played by sensory grounding in getting the system launched in the first place. SH.]

I am willing to concede that COMPLETE grounding is necessary to COMPLETE understanding. But necessity forces us to ignore many details that would provide us with COMPLETE understanding. Complete understanding is unnecessary to living a productive, human-like life, undesirable even. And complete grounding is irrelevant or impossible. An attempt at complete grounding is like trying to satisfy the apparent curiosity of a 3 year-old who has discovered that the single word, "Why?" makes grown-ups talk at length. Eventually, you just end up saying, "Because that's the way it is."

[SH: Grounding is bottom-up, and it's always provisional and approximate, as discussed extensively in "Categorical Perception: The Groundwork of Cognition." The system is always open to revision and updating, at all levels (including the original primitives that got it going in the first place). So "completeness" was never an issue, just sufficiency. SH.]

Are symbol systems inadequate to human-like thought? Heavens, no, they are hallmarks of it! A good deal of the learning process involves learning what to ignore, which is to say, learning what portions of real instances need to be unplugged from their groundings in order to turn them into useful symbols. If one only stopped for red lights which he recognized from previous encounters (and perhaps only from the exact viewpoint from which he was introduced to it), he would make a pretty poor drivers. Ungrounding = generalization = abstraction. The sign of intelligence.

Respectfully submitted, Thomas H. Hildebrandt delta@csl.ncsu.edu

[SH: You have to separate some of your rather hasty interpretations of the grounding proposal from the grounding proposal itself. In the above you have created a little hermeneutic circle of your own. Stevan Harnad.]

Tom, I decided not to post your last two messages to the group as a whole because they are too much in the area of individual speculation and interpretation. In contributions to the group discussion I think it's important to make every effort to pare off our own arbitrary assumptions and start from a point of common understanding and agreement. You have projected extra meanings on unpredicatability, groundedness and completeness that are not shared by everyone (and certainly not implied by what I've written), so I don't think it's profitable for everyone to hear them sorted out. I've responded to you individually this time, because of the effort you put into your won messages, but in future I will not be able to respond to postings that I cannot send to the group as a whole. --- Stevan

----

Date: Wed, 20 Dec 89 01:36:56 EST From: harnad (Stevan Harnad) To: searle@cogsci.berkeley.edu (John Searle) Subject: Re: CR, TT,TTT, SG, and NC

John, you wrote:

> Well Stevan were you surprised by Noam's response? And what
> surprised you? The self regarding interpretation of everything?
> or the incomprehension? or the Stalinoid tone? or something else?
> I am very curious to know.
> Merry Christmas
> John

I was surprised that someone I consider among the deeper thinkers of our time should exhibit such a shallow understanding, even with repeated explication, and I find myself wondering whether the thought process displayed here is in any way representative of its counterparts in his areas of expertise. Probably not, I should think, for I have often enough seen brilliant mathematicians, especially here at the Institute, emphatically expressing the most unrigorous of thoughts on matters outside their domain of expertise (apparently without awareness of their sorry state, rather like an anosognosia). The reason I always find this bewildering is that I have no idea how rationality can be so domain dependent. Creative mathematics, for example, I consider to be perhaps the most intensive use of the human cerebrum: How can that capacity fail so utterly to penetrate ALL of a man's thinking style?

Oh well, I've always been deeply perplexed -- while others have declared it obvious and utterly unperplexing -- about why it is that one can take two people, both having the same knowledge at time T and both having full command of the language, and one can offer them both a string of sentences (an explanation) that leads from their present knowledge state to a new one, and one of them gets there whereas the other does not. Apart from memory lapses or inattentiveness because of lack of interest, I find this common phenomenon utterly mysterious, rather like Achilles and the Tortoise (or the Maine joke whose punchline is: "You can't get there from here!).

Best of the season,

Stevan

Date: Wed, 20 Dec 89 09:51:00 -0800 From: searle@cogsci.berkeley.edu (John R. Searle) To: harnad@clarity

In this case it is a little more complicated: N has spent his life building an edifice. From outside it looks secure but from inside it is desperately insecure. From inside it seems he is always under vicisous attack from people who wilfully pervert his ideas. The world therefore divides into disciples and enemies. Any new idea is a potential threat. Any new idea is only of interest as it relates to his system. If it does not support his system it is probably an attack. Better to deal with it as such than to take chances.

Reread the correspondence with this in mind.

Best

john

From: kirlik%chmsr@gatech.edu (Alex Kirlik)

Stevan Harnad:

Thank you for your thoughtful reply to my objections to your theory. In THE ENTRY POINT PROBLEM, OR WHY/HOW I AM NOT A POSITIVIST, you write:

>There is a vast difference between the project of explicating >scientific theory and of explaining how a robot (like ourselves) >might work. We must be robots before we can be scientists, and >we must explain our robotic capacities before we can explain our >scientific capacities.

Of course I agree with you here. I never meant to suggest that there was any formal similarity between the *problems* of explicating a scientific theory and of explaining how a robot might work. Rather, I intended to point out an important similarity between the two proposed *solutions* to these quite different problems. It is true that I do say that "you are both trying to solve the same problem; namely, how do the symbolic elements in these formal structures get their meanings." But note that the problem mentioned in the previous sentence is only derivative: it is not stated in terms of explicating scientific theories or robotic function. Rather, it is a problem that both you and the positivists have posed for yourself. The reason that you have both been forced to pose this problem is that you have both put forward similar solutions (multi-level symbolic structures capable of of foundational-empirical interpretation) to what I *agree* to be very different problems.

I hope that this clarification of my position lets you conclude that I have not fallen victim to what you call the "entry point problem." I, too, have been a critic of those who commit something similar to the "entry point problem," when they put forward "cognitive models" without any real knowledge of the nature of the information the perceptual system supposedly "supplies" to cognition. Perceptual systems seem to be very smart in supplying rich task-relevant information to cognition, and many of these elaborate cognitive models that presume an impoverished, coarsely-grained, natural language representation (that’s why they have to be so elaborate) raise my ire (in a scientifically dispassionate way, of course).

But I do admit that I am not completely off the hook. You write:

>I don’t care about the "theory-ladenness" of "The snow is white" >as long as the robot can pick out the objects and states of affairs >referred to.

And this sentence made me face one major dissimilarity between yours and the positivists’ enterprise. Namely, the positivists charged themselves with the task of showing how a realist interpretation of scientific theories was indeed possible. Problems of the theory- ladenness of observation dealt them a serious blow in that they could not show how theories could get closer and closer to the "actual truth." But your problem seems only to be to explain how a robot can be successful, so you recognize that there can be "bootstrapping and revision of [perceptual] primitives," i.e., you agree that there need not be one "true" primitive representation of the environment. Only one that works. So yes, I agree that your theory is not susceptible to these "theory-ladeness" arguments that so damaged the positivist position.

And yes, I

>see that there is a world of difference between "observation terms" >-- which are just more symbols -- and the non-symbolic structures >and processes I’m talking about?

But the positivists were, like you, trying to empirically ground meaning in *observation*, not "observational terms." That they hoped for a "physiological theory of perception" to come to their rescue in identifying the "deliverances of the senses" is one indication this is the case. Unfortunately, the positivists assumed the task of explicating the "skeleton" of all scientific theories, and since their chosen language of abstraction was logic, they had to represent the theoretical/observation distinction as a *terminological* distinction; thus, they created "observation terms."

You write:

>Well, I’m inclined to mistrust received views according to which >certain approaches have been decisively discredited unless the >fatal problem with the approach in question is a clearly demonstrable >logical or empirical (or probabilistic) one. (I’ve seen too many >pendula swing back and forth, and too many bandwagons come and go.) >There has been no such clear demonstration of a fatal problem with >certain forms of empiricism.

I agree with your caution; I would guess Minsky & Papert would be a modern lesson. As for the lack of a clear demonstration of the failings of the positivist position, I suspect that this is due as much to the lack of any suitable alternative position, as it is to a perceived lack of problem severity encountered by the positivists’ views. I am sure that positivism would be fully discredited if an acceptable alternative view was available. Van Frassen’s "semantic view" seems to be the closest candidate, as far as I can tell.

I have no futher comments on the rest of your response. Perhaps the symbolist view can even be worked out. Myself, I am happy to agree that people can *be described* as if they were processing symbols, as long as I can add the caveats that a) the vast majority of symbols we process exist in the environment and not in the head; and b) that any symbols in the head somehow manage to process themselves.

Alex Kirlik

From: Stevan Harnad

I could follow all of this until the last paragraph, where you lost me completely. I have tried to describe explicitly what a symbol system is, what the "symbolist view" is, and what’s wrong with the latter (it’s ungrounded). I have no idea what "symbols in the environment" means in this context (as opposed to, say, cultural anthropology or literary theory) and even less of an idea what "symbols in the head processing themselves" might mean (that sounds like hand-waving). I guess that just goes to show why one can’t relegate grounding problems to some unspecified "physiological theory of perception" and then proceed along one’s current ungrounded itinerary, as if it were guaranteed to be independent of and compatible with whatever the "physiological" story might turn out to be. SH.

From: oded%wisdom.weizmann.ac.il@CUNYVM.CUNY.EDU (Oded Maler) To: harnad@Princeton.EDU Subject: On Thinking, Flying and Being

If Thinking is like Flying, then it Cannot be Separated from Being [This started as my summary of the discussion up to 30 Nov 89, but diverged..]

1. Searle’s argument (and its (ir)relevance):

Even if a Turing machine (TM) passes the Turing test (TT), it still does not *understand*, *think* &c.

Reactions:

1) What is this *understanding* that we deny the machine from having, and attribute so easily to ourselves and to other persons?

[SH: It’s what you have (and know exactly what it’s like to have) for English and Hebrew, but not, presumably, for Hungarian or Chinese. That’s all there is to it. SH.]

2) So what? What are the implications?

Counter-reactions:

1) Our own understanding (of, say, sentences in our native language) is a premise to all of the discussions, otherwise we get into loops. The understanding of others who, are (in decreasing order?) members of our culture, of the same age, live in similar geographic conditions,..., humans, mammals, ... , have nervous systems, ..., made of organic materials, .. can be assumed in various degrees of approximation. This is promised to be clarified in yours forthcoming "Other Minds &c", so let’s leave it for the time being.

2) The implication is minor. It just says that an AIer who has written a "story understanding program" doing some parsing and reasoning on texts, and claims that his/her program *understands* stories, is speaking nonsense (or inventing a new meaning for "understanding" which has nothing to do with the common usage). Much of the arguments I’ve seen here consist of attempts of some symbolicists, trying to keep the attribution of understanding from being taken away from their programs, as if it is a necessary condition for their construction to be a noble research goal. I don’t think any reasonable person believes today in the TM-no-robotics version of "strong AI" (perhaps in 1980 this was not so clear, and perhaps I’m not reasonable sometimes). Yet a machine can be very useful without understanding in the human sense. Attributing "knowledge" to machines in the sense of a thermostat is just a conceptual tool that helps in reasoning about their behaviors (cf. knowledge in distributed systems (Halpern and Moses) and knowledge in robotics (Rosenchein and Kealbling)). These notions are much more modest than our knowledge.

[SH: But some of us are not just interested in designing useful machines but in figuring out how the mind works. And some of us (including many who have argued for the other side in this discussion) think thinking is just symbol crunching. SH.]

2. Thinking vs. Flying (I)

Mantal activity at early stages of development is more "behaviorist", linking observable input to observable output in a similar way as an airplane interacts with the environment in order to produce flying. Later more "internal" forms of thought evolve, leading, say, to a mathematician closing his eyes and thinking about A => B, regardless of their groundings. The occurrence of such a thought corresponds presumably to some pattern of internal activity in the brain as experienced by introspection. Here it's hard to find intuitive counterparts from flying.

[SH: If flying is too behavioral, pick heating; and if that's too thermal, pick transduction. And if that sounds too sensory, perhaps that's too bad... They're all nonsymbolic. SH.]

People in academia who usually do more abstract and detatched work, sometimes forget the behavioral aspects of thinking and of speech acts, and jump to strange purely-symbolic conclusions. [SH: It is not at all obvious that abstract thinking must be a symbolic process either. SH.]

3. The Symbolic vs. the *Real Thing*

I think that several important distinctions got mixed in this discussion. In particular:

Discrete / Continuous (= Digital/Analog) Symbolic / Analogical Representation / Reality

Suppose you put a weight on a spring, and the spring reacts accordingly. This is the "real thing". Suppose now you put the same weight on a robotic (or natural) arm. Now the pressure is transduced into an *electrical* (~mathematical, informational) signal, goes into an electrical mechanism that computes some function of the input signal, which goes out, transduced back into muscle\engine activity yielding the external behavior. If you build a general (analog or digital) computing device that calculates the correct (electricity in - electricity out) response function, is it the real thing or a symbolic representation? The mechanism operating on this signal is independent of the signal's "pre-informative" origins (that is whether before being transformed into electricity it was a pressure profile or light intensity or whatever). So your transduction is, in fact, two things - 1) Translating matter/energy into information (and vice versa), and 2) Transforming information from one representation into another (e.g. analog-to-digital).

[SH: It all depends what you're up to. For designing useful devices, these distinctions may not matter much. For mind-modeling and the Total Turing Test (TTT) they're crucial. But recall that I only call for the wherewithal to pass the TTT in order to declare a system duly grounded. Other than that, I simply doubt that the TTT could be successfully paased by just a symbol cruncher married to a few transducers. SH.]

In principle, if one could hook up a kind of "artificial reality" (SciFi, of course) simulation that will perform the simulation of the physical-to-informational part of the transduction (not the analog-to digital), letting every input port (=sensory surface) observe the appropriate electrical patterns, then the rest of the system (doing analog and digital computation) will do the "thinking" (A brain-in-the-vat with simulated external stimuli). [SH: Simulation is simulation, that's all. There is no "artificial reality," just symbol crunching that is interpretable as representing reality. The rest is the hermeneutic hall of mirrors generated by forgetting that you're just projecting an interpretation on a systematic pattern of squiggles and squoggles. I'm not sure what you mean by the brain in a vat and the analog and digital computations: Are you talking about a symbol cruncher or not? If not, all bets are off. (By the way, you may find that making a symbolic model of the world as it impinges on a person for a lifetime -- a "TTT" of his input -- will be much harder than modeling the person himself.) SH.]

I wonder how a system that has a standard computational part augmented with the capabilities of performing the informational part of the transduction (2) will be called: you will say that it’s not a pure symbol system, because of explicitness and systematic interpretation (I’m still not sure about the color of the explicitness herring, but never mind, I guess you derived your definitions of a purely-symbolic system from writings and claims of honorable and respectable men...) Although it does not meet all your criteria, there is some symbolic (or representational, if you prefer) flavour (as opposed to the real thing), after all, the rules of multiplying numbers (by an analog computer), or integrating waves are independent of the physical entities they represent. So maybe we can call it a symbolic-subsymbolic dynamical information processing system. Now, if you claim that such a system does not think, that is, that thinking is like flying, you come to the conclusion that THINKING (in your sense) cannot be separated from BEING, that our thinking is not just information processing, and cannot be captured by some abstraction of the behavior of our nervous systems.

[SH: Not at all. Thinking may be lots of kinds of processes, but not just symbol manipulation. I’m still not sure what sort of processes you’re imagining here, but if they are not just symbolic, I can raise no objection (based on any of the symbol grounding considerations that are under discussion here). SH.]

Before you burst in I’ll try to repeat the argument. You said that simulated planes don’t fly. Why? Because flying is not just an information processing activity nor a subjective feeling based on introspection made by an externally-unobservable dynamical system. It is a well-defined observable physical phenomenon. When we come to "thinking", there are two choices. One is based on observation, i.e., the TT or the Turbo TT ("by their fruits.."). The other is trying to reproduce the subjective feeling. Now, I claim, that if this subjective feeling cannot be reproduced (in principle) by an information processing device coupled with a simulation of the external world (including the physical-to-information part of the transduction, i.e. the rest of the body) that puts at the sensory surfaces exactly what the world would have put, and at the same rate then this sense of thinking cannot be implemented by any non-human device, since it involves other aspects of human existence beyond information processing (whose aspects have already been captured by our hypothetical device).

[SH: Nothing of the sort. As long as your artificial brain in a vat is not just a symbol cruncher, and as long as it could pass the TTT if you gave it the sensors and effectors, I don’t care if you simulate its input (if you can). -- By the way, the only role that "being" plays in this argument is that you have to be the candidate to know for sure that it has a mind. (That is why Searle’s Argument works, and answer "No!" in the Chinese Room.) SH.]

This also reminds me of your question to the connectionists. In the subjective sense of thinking, conciousness etc., there is a difference between a parallel implementation and sequential simulation. If one believes that conciousness is the product of some patterns of activity in space-time, then sequential simulation cannot simulate this subjective feeling in the same (or at least similar) sense that drinking the coffee after eating the sugar do not produce the same feeling as drinking them together. --Oded

[SH: Yes indeed, if you hypothesize that thinking is essentially parallel, then it follows that a nonparallel process would not be thinking. So what? (I would also discourage picking your favorite candidate process on the basis of some analogy to subjectivity. It’s a safe bet that subjective experiences and objective processes are incommensurable: An experience that seems simultaneous may be serial, and vice versa. Stevan Harnad.]

> Date: Mon, 25 Dec 89 21:32:26 EST
> From: Ted Adelson
>
> By the way, I understand that you are the central switching station
> for Chinese Room discussions. I have not paid attention to the debate
> since it first began, but I just saw the articles in Scientific
> American, and was inspired with an interesting thought. Can you tell
> me if this idea has been put forward before?
>
> The Chinese Searle Problem:
>
> As Searle notes, the man can actually serve as the Chinese Room
> himself by memorizing all the instructions. Then he will appear to
> understand Chinese. Now suppose that the man is Searle himself.
> Further, suppose that we construct the instructions so the the Chinese
> person acts as Searle would act if he were Chinese. We interrogate
> Searle in English and he says, "Well, I appear to speak Chinese, but I
> actually have no idea what any of it means -- I’m just executing the
> instructions mechanically." We then interrogate Searle
> in Chinese and he says, "Well, I appear to speak English, but I
> actually have no idea what any of it means -- I’m just executing some
> instructions mechanically." Both the Chinese Searle and the English
> Searle behave in exactly the same way, and express the same opinions
> about the sham they are executing in the foreign language.
>
> I would argue that we now have two minds that happen to coexist in the
> same brain -- i.e. they are both instantiated in the same piece of
> hardware. The English Searle really doesn’t understand Chinese, just
> as he says, and the Chinese Searle really doesn’t understand English,
> just as he says. But the Chinese Searle really does understand Chinese.
>
> The English Searle might point out that the Chinese word for "chair"
> does not excite brain states that have anything in common with the
> states generated by the English word for "chair." This is correct, but it
> simply proves that the English Searle doesn’t understand Chinese,
> which we already agreed was true. The point is that the Chinese
> Searle does make all the right associations to the Chinese word for
> "chair," and so he does understand its meaning.
>
> Please -- I don’t want to get swamped with a complete list of all

> Chinese Room arguments, and I don't want to become a part of the
> ongoing electronic debate. But could you just give me a quick comment
> on the above line of thought?

Yes, this kind of point has been brought up and it doesn't work. First, it would be trivial to add to a (hypothetical, perhaps counterfactual) program that already could pass the Turing Test -- i.e., could babble with you for a lifetime just as a real pen-pal would, so you could never tell the difference between it and a real pen-pal -- the extra feature that, if you happened to ask it whether it spoke English it would say "No, I just memorized a bunch of instructions for manipulating meaningless symbols and I'm just going through the motions." That adds absolutely nothing, logically or methodologically, to the issue at hand. It would still be true that Searle HAD memorized the instructions, and really did understand English and not Chinese. The computer program would contain instructions for generating symbols strings that were likewise interpretable as claiming it had done what Searle had done, and did/didn't understand what Searle didn't/did, though it hadn't, and that would just be another piece of empty mimickry -- going through the symbolic motions in a way that is INTERPRETABLE by us as if it were understanding, but with no real understanding going on at all (apart from the understanding going on in and being projected from the heads of us human interpreters).

Look, the way your point is usually put (without needlessly iterating, by adding the extra little wrinkle you propose, the hall of mirrors that has already been created by our projecting meanings onto the symbols) is that in the Chinese Room Searle simply has two minds in his head: One that understands English but not Chinese, and another that understands Chinese (but not English, if you wish), and the two minds are inaccessible to one another, as in multiple personality disorder.

The quick way to ween oneself from sci fi fantasies like this (unless one is already hopelessly lost in the hermeneutic hall of mirrors) is to try to keep reminding oneself that there is no evidence WHATSOEVER that memorizing a bunch of instructions for manipulating squiggles and squoggles is a possible cause of multiple personality disorder! (You need stronger stuff, like early child abuse...)

Ted, I'm glad that Searle's article has stimulated some thoughts for you, and was happy to reply "in camera" this one time. But the fact is that there are some deep issues here, about which I have been doing quite a bit of thinking for some time. I can by now see quite clearly the errors in the dozen or so standard attempts to rebut Searle (of which yours was one or two) that keep resurfacing over and over. I'm willing to take the time to discuss the more interesting incarnations of them on the net, even if there is evidence that my interlocutor has not given it nearly as much thought as I have, but I can't do it on a one-to-one basis each time because there just isn't time. (Perhaps I'll create a little "Chinese Room" of my own -- it wouldn't have to be big, just about 12 counterarguments written out long-hand -- and simply distribute it automatically to all new comers who come bearing familiar old nonstarters...)

Cheers, Stevan

Summary: Even if no coherent dual interpretation of a symbol system is possible, it's all in the mind of the interpreter Date: 10 Jan 90 16:14:48 GMT kck@g.gp.cs.cmu.edu (Karl Kluge) of Carnegie-Mellon University wrote:

> Are we to suppose that you can impose an interpretation on the symbols
> such that they form both a coherent conversation in Chinese *and* a
> legal and coherent chess game?

Your point about there not necessarily being more than one coherent interpretation of a symbol system is correct. (In fact, I've made the point myself in print, about both the alleged "radical underdetermination" of language translation (Quine/Goodman) -- in which it is supposed, without proof, that meanings can be swapped willy nilly while preserving a coherent overall semantic interpretation -- and about "spectrum inversion" -- in which subjective quality is imagined to be similarly swappable, say, green looking red and vice versa, again on the assumption that the psychophysics, discourse and behavior could be coherently preserved under the transformation.)

To assume that such swaps are possible (always, sometimes, or even ever) is equivalent to assuming that there exist semantic (or behavioral) "duals," very much like mathematical duals such as not/and vs. not/or in the propositional calculus. There is neither proof nor evidence to support such an assumption except in particular formal cases, such as the nonstandard interpretations of arithmetic.

However, Searle's point does not depend on this assumption! He's not claiming that the alternative interpretations would be COHERENT; he's only reminding us that projecting an interpretation is clearly all that's involved -- in either the coherent or the incoherent case. The difference is subtle, but crucial to understanding Searle's (perfectly valid) point.

Stevan Harnad

To: Symbol Grounding Discussion Group Please let me know if you wish your name removed form this list. Stevan Harnad

----

Comments from (and replies to): Massaro, Stede, Mocsny, Yeap and McDermott SYNTAX, SEMANTICS, and HYBRID MODELS

> From: psych36@ucscd.UCSC.EDU (Dominic Massaro)
> To: Stevan Harnad
> Subject: symbol grounding
>
> I found your grounding paper very informative, but I am puzzled about
> one thing. Your fourth property of a symbol system precludes "meaning"
> of the tokens playing a role in the manipulation of these tokens by
> rules. I don't see why this constraint is necessary and it seems
> wrong. If we are categorizing letters--say in word recognition--,
> neighboring letters can influence this categorization and not just the
> internalized shape of the letter.
>
> More importantly, your constraining properties of symbol systems seem

> to make a hybrid model impossible. The connectionist front-end would
> putatively give meaning to the symbols which would be functional in the
> syntactic operations.
>
> I don’t see how you can have it both ways. Dom

I’m afraid you’ve misunderstood. The syntactic/semantic distinction is a given here. Of course meaning is involved in letter recognition, but that just means letter recognition is not being done just syntactically, i.e., not just by symbol manipulation. The symbol grounding problem is a problem only for a purely syntactic approach to cognition.

But note that you can’t help yourself to semantics for free either. You have to say in what it consists. The syntactic approach bravely tried to claim that it consists in doing the right symbol manipulations. It turned out to be wrong. To get into the game you must propose a rival.

I do advocate a hybrid model, as described in the paper. There are (1) analog copies of the sensory projections and (2) categorical representations that pick out sensory categories (and allow them to be named) by selecting only the invariant information in the sensory projection. Finally, (3) the names of the sensory categories are the grounded elementary symbols and the rest of the symbol system is composed out of these. Connectionism is a candidate mechanism for learning the invariants. The system is hybrid nonsymbolic/symbolic because the analog and categorical representations are nonsymbolic -- they are not based on syntactic manipulation of arbitrary shapes but analog and causal relations to the objects they are projections of.

Stevan Harnad

EXPLANATION AND CATEGORIZATION

> From: stede@cs.purdue.edu (Manfred Stede)
>
> After returning from the vacation I had time to look at the symbol grounding
> paper. I’m doing a Master’s degree in CS and have studied Linguistics and (a
> bit) Philosophy as well. After reading a lot about semantics I felt quite
> uncomfortable because people tend to circumvent the question of how words
> finally acquire a ’meaning’, apart from being defined in terms of other
> words. I.e., exactly the point you address.
>
> Let me make just a few (rather spontaneous) comments on the paper. On page
> 3 you seem to equate generating intelligent behavior with explaining it (the
> parenthesis ’hence explaining’) - isn’t this a step back to behaviorism? If
> I write a program that somehow plays chess, did I explain how humans play
> chess? - To my mind, it is the objective of "weak AI" to produce intelligent-
> like behavior, whereas "strong AI" goes ahead and claims the program to
> *explain* something with psychological validity.

Behaviorism never explained how you generate behavior, internally, just how you shape it, externally. If you can generate a behavior, you can explain at least ONE way it can indeed be generated. Whether it’s the way our brains (or equivalent systems) do it is another question. I point out in this and other papers that when the behavior is of the scale of our total robotic capacity (the Total Turing Test, the TTT), including, if necessary, even all of our neurons’ "behavior" (the "TTTT"), then the problem of whether the explanation is "right" is identical with the problem of underdetermination for any theory, including physics. The only extra degree of freedom is the question of whether or not the device is conscious. This I think you can only know if you ARE the device; otherwise, it’s an insoluble problem, unique to the theory of mind. But, being an epiphenomenalist, I don’t think anything of causal consequence rides on it.

> I liked your explanation of the chinese room. I think it is somewhat clearer
> than Searle’s own words.
>
> Do you believe in clearly bounded categories? I feel that Wittgenstein’s
> illustration of "game" applies to a whole lot of other words/concepts. I’m
> highly skeptical about the value of any "absolute judgement". How about an
> ordinary chair which is deprived of one of its legs - is it still a chair?
> Closely related: Are there really *invariant features* to every category?
> Again, Wittgenstein showed convincingly that for "game" this is not the case.
> I guess the point is that your analysis is restricted to perceivable objects
> of the real world, where things might be somewhat more straightforward than
> with abstract entities like "creativity" etc. Anyway, I’m not sure whether in
> the long run one should distinguish between real-world entities and abstracts
> - meaning is made in heads and only there; thus I feel that a semantic theory
> should account for boths classes of concepts in the same way.
> (But I admittedly don’t know how...)
>
> So much for the moment, please keep me on your mailing list, Manfred Stede

I have an argument in that paper, and in Categorical Perception: The Groundwork of Cognition, against what Wittgenstein "showed convincingly." In brief, there are determinate all-or-none categories, like chair and bird as well as graded ones, like big. There are instances or regions of variation in the former (like your legless chair) that are indeterminate or matters of degree, like the latter. There are also stipulated categories, like game, where it’s arbitrary what you choose to take as a game, or you don’t know, or you can’t know. So what? For a categorization theorist (and a grounding theory) you need only explain the determinate sorting that we CAN successfully accomplish, and that must have an invariant basis, otherwise how do we do it?

Stevan Harnad

UNDERSTANDING VERSUS LEARNING TO UNDERSTAND

> Date: Sat, 6 Jan 90 11:48:49 EST
> From: daniel mocsny
>
> Steven,

>
> Thank you for your reply.
>
> > It’s irrelevant whether or not Searle would eventually learn Chinese
> > from manipulating symbols. The hypothesis was that the understanding
> > consists in the manipulation of symbols.
> >
> > Stevan Harnad
>
> I agree that’s true at one level; Searle does not "understand" while
> he is only manipulating rules one at a time without having yet
> "internalized" them. However, Searle quite clearly asserts that he
> can’t *ever* establish semantics from sitting in the Chinese Room,
> regardless of how long he stays in there. And in your post, you stated
> that if Searle memorizes the rules nothing changes. I think I have to
> disagree with both of those statements.
>
> I strongly suspect that a human mind can (and invariably does) impose
> a consistent semantics on purely syntactic input after enough practice,
> and I believe this is relevant to the discussion. I see two Chinese
> Rooms---one containing an initially ignorant novice, and the second
> containing a seasoned symbol pusher. The first Chinese Room clearly
> does not contain anyone who understands, and a computer that simply
> models what Searle is doing in there will not understand either.

You may as well stop right here, because the latter is all that a computer does. You cannot project onto it Searle’s eventual capacity to figure out what the symbols might mean (that’s completely irrelevant). Your projection is yet another symptom of the "hermeneutic hall of mirrors" I keep warning about.

> But the second Chinese Room contains someone who does understand
> in some sense (and I think that understanding can become arbitrarily
> close to the "real thing"). And Searle has not even acknowledged that,
> much less demonstrated why a computer could not be programmed to do
> the same thing.

The second Chinese room contains a person, who may be able to figure it out eventually, using, among other things, his knowledge of a grounded first language. That’s completely irrelevant to the computer in the first room, whose "first" language is supposed to be Chinese, whereas Searle has shown it’s nothing of the sort.

> (Note: I have no idea whether a computer could be programmed to do the
> "same" thing. I’m sure a computer could simulate it, provided we
> understood the mechanisms of skill acquisition in humans. Searle might
> still argue that the system doesn’t "really" understand, but the
> distinction is no longer at all so clear-cut.)

The distinction is as clearcut as it has been all along. If you don't inadvertently and unwarrantedly project 2nd order human capacities on the computer, like Searle's capacity to learn Chinese as a second language (perhaps even to decrypt it eventually in the Chinese room), then it will be clear that there is only ONE Chinese room problem. Your 2nd order problem is the same as the 1st order problem. Searle could show that a computer mimicking him in your second room was as empty of understanding as the first.

> In other words, the Strong AI problem is really a two-level problem,
> and early AI systems (as well as Searle's Chinese Room argument) focus
> only on the first part of the problem, where understanding does not
> exist. That is, a system that works by explicitly pushing symbols that
> correspond directly to logical rules mimics an aspect of human
> intelligence that does not contain everything necessary to constitute
> "understanding". In other words, rule-based systems imitate some of
> the behavior that normally proceeds from understanding, without
> modeling the understanding itself. Sort of like building a clay model
> of a locomotive and wondering why it doesn't run...But the clay
> model's failure in no way precludes the eventual success of a model
> that incorporates the essential structure of the real locomotive.

Yes, but that model will no longer me just clay; and the model that passes the TTT will not be just symbolic. Which is the point of the symbol grounding argument.

Stevan Harnad

AI, MIND-MODELING and SYMBOLS

> To: Stevan harnad
> From: W.K. Yeap COSCWKY@otago.ac.nz
>
> I thought I better drop you my comments without further delay. I
> haven't got the time to digest the many replies that you had right now
> and so my reply is directly on your paper. I apologise if any of these
> remarks have been made before. I am off to a symposium in Stanford and
> will read through the others when I get back.
>
> Unfair to AI
>
> I am a strong believer that a major (& real) AI problem NOW is the symbol
> grounding problem as you have described. In this respect, I particularly
> like the way in which the problem is formulated in your paper.
>
> >From an AI standpoint, I believe our contribution is to show how -
> (at least demonstrate one possible approach - using the computers we
> have) - both in terms of identifying and describing the processes
> involved and the representations required. I am of course advocating
> Marr's approach to AI.

>
> This leads to my first (albeit a minor) complaint on your paper which
> is when you claim that the symbol grounding is a problem only for cognitive
> modeling, not for AI in general, even though you ended your note by
> saying that "even in AI there may be performance gains to be made
> (especially in robotics and machine vision) from endeavouring to
> ground symbol systems". I disagree and conclude below that the
> symbol grounding problem is a major AI problem. (There is only
> one AI - a strong AI). Although many people out there are creating
> "AI" systems (esp. commercial systems) their work does not contribute
> much towards the building of intelligent machines (cf. that satisfy
> the kind of criticisms made by Searle) and should really be disregarded
> as AI work in the sense that programs like Eliza and programs
> for playing chess are no longer AIish. The idea of generating
> complex behavior through symbol manipulation must be firmly
> established by now and by simply trying to make one more program "smarter"
> than another is not AI. It is no wonder that we now have many
> subfields of AI, the problem tackled in most cases (exception being
> vision) is really a problem in the subfield domain and not AI. >
> A consequence of the above view is that (the real concern of) AI
> research should concentrate on building processes which are based on
> inputs eventually derived from the perceptual end. (Note that I
> do not imply that all AI researchers must work on, say, low-level
> vision or categorical perception but in their chosen domain,
> considerations must be given as to how their representations
> could be derived from our perceptual systems i.e the grounding
> problem! The more vague the connection is, the more effort should
> be spent working on it than on anything else). One well-known example
> of such AI work is Marr's vision work.

It seems to be perfectly legitimate to try to make machines do intelligent and useful things without caring about whether they are doing it in a mind-like way. The symbol grounding problem affects such work only to the extent that grounding may be needed to get the job done at all.

> And now I have my second complain about the paper.
> You pointed out that one weakness of the standard reply of the
> symbolist (cf sect. 3.3) is that .. "it trivializes the difficulty
> of picking out the objects, events and states of affairs..."
> Yet in your proposed solution (cf sect. 5), you did almost the
> same when saying "our ability to discriminate inputs depends on
> our forming iconic representations of them. These are internal
> analog transforms of the ..." What analog transforms??? We have
> a glimpse of the complexity of these processes in AI research on
> early vision (Marr, Poggio, ...). You do no justice to AI researchers
> working in this area by saying we need horse icons to discriminate
> horses. The complexity involved in obtaining the horse icons from
> the intensity images surely make it meaningless to say that "icons of
> sensory projections are too unselective". Worst still, in section 6,

> you seem to address this problem by suggesting that a solution may be
> afforded by people working in connectionism. It is unfair that you
> make no mention of the efforts of the AI researchers, esp. those
> working on early vision. One possible reason could be that you
> consider such work as not symbolic (cf 5.3) & therefore traditional
> AI approach is irrelevant.

I have the utmost respect for both robotic and early vision (some of which is symbolic, some not). At the level of generality at which I was discussing these problems in the paper, the particulars of, say, Marr's approach were simply not relevant.

Connectionism was only proposed as one natural candidate for the learning component of the grounding scheme I described.

> This leads to my closing remark.
> IF we are trying to understand intelligent systems as symbol
> manipulating devices, the symbol grounding problem
> is indeed the next major problem FOR demonstrating that such systems
> can understand, without being just a system in a Chinese Room. THE
> PROBLEM NOW IS TO DISCOVER HOW and it does not matter whether we
> pose the problem as an AI or Cognitive Modeling or Connectionism.
> The other hallmark of AI is that the ideas are implemented and
> detailed computational processes are made explicit & for this
> reason, I like to call it an AI problem (there is no worry that the
> solution may incorporate connectionists ideas!). It is a pity to see
> so few AI projects reported in the journals have any relevant to
> this problem! Let us work on it!

The question my paper (and Searle's) raised was whether standard, purely symbolic AI -- Strong AI -- could ever succeed in generating a mind. I gave reasons why it could not.

Stevan Harnad

IS THERE AN OBJECTIVE-SUBJECTIVE CONTINUUM?

Drew McDermott (mcdermott@cs.yale.edu) wrote:

> Now, on the question of whether awareness is an all-or-nothing thing:
> You argue that there is a strict dividing line between being aware and
> not being aware. If I had to guess, I would guess there is no such
> dividing line.

If you have followed what I said about being aware of more or less (which is ok) or being more or less aware (which is also ok) versus having no awareness at all, then I challenge you to give a coherent notion, even, of what the nonaware-aware continuum might be (much less what it might be like -- since the notion seems to call for a continuum with respect to whether something is "like anything" at all).

What's at issue is like the difference between zero and nonzero: A quantity is not not more or less nonzero, it's either nonzero or it isn't! Both logically and phenomenologically, I think, the kind of continuum you have in mind is simply incoherent. Not so for a continuum between inert and living, whose continuity is in no way problematic, since only objective properties are at issue. But what a nonawareness-to-awareness continuum requires is a graded passage from objectivity to subjectivity, and that's what I claim is incoherent.

What is the intuitive source of this incoherent notion? You already mentioned it: A spurious analogy with degrees of awakeness (sic) in an already known-to-be-aware subject: ourselves.

> But what about the methodological questions? A computationalist could
> say, Sooner or later we'll know by inspection of different brains the
> degree to which any of them are aware. A "subjectivist" would have to
> say, We can *never* really know what it's like to be a bee (say), so we
> can never know whether bees are best described as "half-aware," so we
> can never know whether there is such a thing as half awareness. I would
> have thought you were in this camp, and would remain agnostic on the
> question of whether awareness could be partial. -- Drew McDermott

I am in the camp that recognizes that the other-minds problem is insoluble: The "methodological" question of whether or not anyone but me is aware is simply unanswerable. Now I've bitten the bullet and accepted that the TTT (possibly augmented to the "TTTT," to include the observable behavior of my neurons) is the most that an empirical theory of mind will ever have by way of objective data, and so I'm prepared to believe that whatever can pass the TTT(T) (i.e., can behave completely indistinguishably from one of us) is aware.

But there's nothing about "degree of awareness" in that. The TTT applies to people. We don't know (and I don't think we'll ever know) the ecology of any other organism completely enough (or intimately enough, for we also lack the introspective empathy to be able to judge, as we can with out fellow humans, what nonhuman mental states might be like) to administer a nonhuman TTT with any confidence, so that's no source of degrees of awareness either.

The observed correlates and causal substrates of TTT-passing capacity may turn out to admit of degrees, but nothing at all follows about degrees of awareness from that either. So there you have it.

(And there are of course other alternatives to "subjectivism" besides computationalism; in fact, that's what's at issue in the symbol grounding discussion...)

Stevan Harnad

Alexis Manaster-Ramer AMR@ibm.com wrote:

> I know of no... attempt [to make sense of "content" or
> "intentionality"] that does not either (a) work equally well for robots
> as it does for human beings (e.g., Harnad's proposals about
> "grounding")... [G]rounding as I understand Harnad's proposal is
> not... intended to [to distinguish biological systems

> from robots]. I think he contends that robots are just as grounded as
> people are, but that disembodied (non-robotic) programs are not.

You're quite right that the grounding proposal (in "The Symbol Grounding Problem," Physica D 1990, in press) does not distinguish robots from biological systems -- because biological systems ARE robots of a special kind. That's why I've called this position "robotic functionalism" (in opposition to "symbolic functionalism").

But you leave out a crucial distinction that I DO make, over and over: that between ordinary, dumb robots, and those that have the capacity to pass the Total Turing Test [TTT] (i.e., perform and behave in the world for a lifetime indistinguishably from the way we do). Grounding is trivial without TTT-power. And the difference is like night and day. (And being a methodological epiphenomenalist, I think that's about as much as you can say about "content" or "intentionality.")

> Given what we know from the theory of computation, even though
> grounding is necessary, it does not follow that there is any useful
> theoretical difference between a program simulating the interaction
> of a being with an environment and a robot interacting with a real
> environment.

As explained quite explicitly in "Minds, Machines and Searle" (J. Exp. Theor. A.I. 1(1), 1989), there is indeed no "theoretical difference," in that all the INFORMATION is there in a simulation, but there is another difference (and this applies only to TTT-scale robots), one that doesn't seem to be given full justice by calling it merely a "practical" difference, namely, that simulated minds can no more think than simulated planes can fly or simulated fires can burn.

And don't forget that TTT-scale simulations have to contend with the problem of encoding all the possible real-world contingencies a TTT-scale robot would be able to handle, and how; a lot to pack into a pure symbol cruncher...

Stevan Harnad

TT, TTT, TTTT and Stronger...

Alexis Manaster-Ramer (AMR@IBM.COM) wrote:

> (1) I am glad that SOME issues are getting clarified, to wit, I hope
> that everybody who has been confusing Harnad's arguments about
> grounding with other people's (perhaps Searle or Lakoff's) arguments
> about content, intentionality, and/or semantics will finally accept
> that there is a difference.
>
> (2) I omitted to refer to Harnad's distinction between ordinary robots,
> ones that fail the Total Turing Test, and the theoretical ones that
> do pass the TTT, for two unrelated reasons. One was that I was not trying
> to present a complete account, merely to raise certain issues, clarify
> certain points, and answer certain objections that had arisen. The other
> was that I do not agree with Harnad on this issue, and that for a number

> of reasons. First, I believe that a Searlean argument is still possible
> even for a robot that passes the TTT.

I'd like to hear that Argument! As I think I showed in "Minds, Machines and Searle," the robotic version of the Test is immune to Searle's Argument, which works only against pure symbol manipulation. The reason it works there is both symbol-manipulation's chief weakness and its chief strength. "Symbolic functionalists" ("Strong AI") call it "multiple realizability" and I lampoon it as "teleporability": It's basically the software/hardware distinction. Here's the argument, step by step:

The same program can be realized on many different hardwares. What matters is that in every implementation of the program the internal states stand in a certain functional relation to one another (and to inputs and outputs). Hence, any functional property a given formal program is claimed to have, ALL of its possible implementations must have that property too. So, in particular, if it is claimed that, say, understanding Chinese, is just symbol manipulation of the right kind, then once you commit yourself to the program in question (one that passes the Turing Test [TT] -- symbols-only version), EVERY IMPLEMENTATION OF IT MUST ALSO BE UNDERSTANDING CHINESE ("teleportability"). So, whereas the "other minds problem" normally presents an impenetrable barrier to ascertaining whether any other system than oneself is really thinking, in this special case, Searle himself can BE the implementation -- and can accordingly come back and report that, no, he did not understand Chinese in so doing, hence neither could that pure symbol-crunching computer (to which he is functionally equivalent, remember) have been understanding Chinese.

That's why the Argument works for the TT and pure symbolic functionalism. But, as I showed with my Transducer Counterargument (where the mental function under scrutiny happens to be, instead of understanding, seeing), it doesn't work for seeing, because seeing depends ESSENTIALLY on (at least one) nonsymbolic function, namely, transduction; and transduction, because it's nonsymbolic, either cannot be implemented by Searle (in which case the Chinese Room Argument cannot even be formulated, for Searle cannot duplicate all of the robot's functions, on pain of a VALID "Systems Reply," for a change) or he must BE the robot's transducers -- in which he case he would indeed be seeing. Now sensing and acting on the objects in the world are robotic functions. They are required by the TTT, but not the TT (at least not directly). Hence the TTT is immune to the Searlean argument.

> Two, the TTT is much too strong,
> since no one human being can pass it for another, and we would not be
> surprised I think to find an intelligent species of Martians or what
> have you that would, obviously, fail abysmally on the TTT but might pass
> a suitable version of the ordinary Turing Test.

Pass it for another? What does that mean? The TTT requires that a candidate be completely indistinguishable FROM a person, TO a person, in all human robotic and linguistic capacities.

You seem to be misunderstanding the logic and the methodology of the TTT. The TTT applies only to our species, and other candidates from which we cannot be distinguished. It is based on the simple intuition that, whatever it is that convinces us that anyone other than ourselves really has a mind, it can't really be anything we know about how the mind works internally (since we don't know how the mind works internally). Hence if a candidate acts (for a lifetime) indistinguishably from a real person with a mind, like ourselves, then we are in no logical or methodological position to deny that he has a mind because of any arbitrary notion we have about how he works (or should work) internally. What distinguishes the TT from the TTT, however, is that, be it ever so convincing, a lifelong penpal test is not strong enough, and not really the one we use with one another. We ask for indistinguishability in all interactions with the world (including robotic ones).

Other species, including extraterrestrial ones, are simply irrelevant to this, both logically and methodologically. In point of fact, we have trouble Turing Testing other species because we don't have the intuitions about what it's like to BE them that we do have for our own species. Nor can we ever know their ecology well enough to know whether they are Turing indistinguishable from their OWN kind. (For similar reasons I will only bring up if someone asks, congenitally deaf/dumb/blind people, paraplegics, etc. are not relevant to this discussion either.)

> Third, the TTT is still
> a criterion of equivalence that is based exclusively on I/O, and I keep
> pointing out that that is not the right basis for judging whether two
> systems are equivalent (I won't belabor this last point, because that is
> the main thing that I have to say that is new, and I would hope to address
> it in detail in the near future, assuming there is interest in it out
> there.)

For reasons already discussed in this symbol grounding discussion group, I don't find your arguments for deeper equivalence convincing. There is, of course, a logical possibility that of two TTT-passing robots, one has a mind and the other doesn't. How could we ever know? Because one of them is more brain-like in its inner functions? How do we know how much brain-likeness is necessary or sufficient? And if insufficient, how do we confirm that it did not, indeed, succeed in generating a mind, despite the TTT-power? It seems as if your stronger functional equivalence criterion is an a priori one, and arbitrary, as far as I can see, for there is no way to determine whether or not it is really necessary.

Of course, it's always safer to ask for more rather than less. What I've called the "TTTT" requires indistinguishability not just in robotic performance capacity (what our bodies can do) but in neural performance capacity (what our cells and even our biomolecules can do). That's as much as empiricism and objectivity can furnish. I'm betting, however, that this extra fine-tuning constraint will be unnecessary: That once the problem of generating TTT performance itself is solved (no mean feat, depite how easy it is to suppose it accomplished in thought experiments like Strong AI's and Searle's), ALL the successful implementations will have minds (I call this "robotic functionalism"), and that aiming at generating TTT-capacity will be a strong enough constraint to guide our research.

What's sure is that we can never know the difference (except in the special teleportable case of symbol crunching). This is one of the reasons I am a methodological epiphenomenalist.

> (3) Likewise, I omitted to refer to Harnad's position on simulation because
> (a) I thought I could get away with it and (b) because I do not agree with
> that one either. The reason I disagree is that I regard simulation of a
> system X by a system Y as a situation in which system Y is VIEWED by an
> investigator as sufficiently like X with respect to a certain (usually
> very specific and limited) characteristic to be a useful model of X. In
> other words, the simulation is something which in no sense does what

> the original thing does. However, a hypothetical program (like the one
> presupposed by Searle in his Chinese room argument) that uses Chinese
> like a native speaker to engage in a conversation that its interlocutor
> finds meaningful and satisfying would be doing more than simulating the
> linguistic and conversational abilities of a human Chinese speaker; it
> would actually be duplicating these. In addition--and perhaps this
> is even more important--the use of the term simulation with respect to
> an observable, external behavior (I/O behavior again) is one thing,
> its use with reference to nonobservable stuff like thought, feeling,
> or intelligence is quite another. Thus, we know what it would mean
> to duplicate (i.e., simulate to perfection) the use of a human language;
> we do not know what it would mean to duplicate (or even simulate partially)
> something that is not observable like thought or intelligence or feeling.
> That in fact is precisely the open question.

"Viewing-as" is just hermeneutics. It is not a matter of interpretation whether I, as a "model" of someone who understands English, am really understanding English. I either am or I am not. Searle's special little privileged periscope is here to tell us that implemented symbol crunchers do not. That normally unobservable stuff is, after all, observable to ONE observer, namely, the subject -- if it's there, that is. And Searle is in a position to report that, in a symbol cruncher, it's not. (This is the classical point where people jump in with irrelevant non sequiturs about unconscious understanding, etc. etc. I'll reply only if someone -- sigh -- insists on bringing it up.)

The conclusion, by the way, is not just that symbol crunching is NOT understanding, but also that the TT is insufficient, for at least one kind of candidate could hypothetically pass it while demonstrably not understanding. Hence linguistic performance is no more an isolable mental module than any other fragment of our performance capacity (like chess-playing, theorem-proving or block-manipulation) is. Nothing less than the TTT will do; the rest are just underdetermined (and mindless) toys.

> And, again, it seems to
> me that the relevant issue here is what notion of equivalence we employ.
> In a nutshell, the point is that everybody (incl. Harnad) seems to be
> operating with notions of equivalence that are based on I/O behavior
> even though everybody would, I hope, agree that the phenomenon we
> call intelligence (likewise thought, feeling, consciousness) are NOT
> definable in I/O terms. That is, I am assuming here that "everybody"
> has accepted the implications of Searle's argument at least to the
> extent that IF A PROGRAM BEHAVES LIKE A HUMAN BEING, IT NEED NOT
> FOLLOW THAT IT THINKS, FEELS, ETC., LIKE ONE. Searle, of course,
> goes further (without I think any justification) to contend that IF A
> A PROGRAM BEHAVES LIKE A HUMAN BEING, IT IS NOT POSSIBLE THAT IT
> THINKS, FEELS, ETC., LIKE ONE. The question that no one has been
> able to answer though is, if the two behave the same, in what sense
> are they not equivalent, and that, of course, is where we need to
> insist that we are no longer talking about I/O equivalence. This
> is, of course, where Turing (working in the heyday of behaviorism)
> made his mistake in proposing the Turing Test.

(1) Your "strong equivalence" criterion is at best a metaphysical insurance policy you can never cash in; at worst it's arbitrary. Performance capacity is the only real arbiter.

(2) No one wants to DEFINE mental states in I/O terms. (Who's in any position to do any defining!) Empirical methodology is the only thing at issue here. You can never know that a stone hasn't got a mind, of course. And even Searle can't prove that that computer over there can't understand even though when he does everything it is doing internally, he doesn't understand. He can't even prove he hasn't got TWO minds, one the regular one, and the other a result of multiple personality disorder engendered by the memorization of too many symbols. So "necessity" and "possibility" are not at issue here; that's too strong.

(3) The only thing Turing got wrong, in my view, was the systematic ambiguity at the root of the difference between the TT and the TTT. And Searle has succeeded in showing (by a purely intuitive argument) that by the usual probabilistic canons of empiricism, it is highly unlikely that either he or that computer understands Chinese. A need for some "stronger" form of equivalence has absolutely nothing to do with it.

Stevan Harnad

ON PLANES AND BRAINS

Donald A Norman norman@cogsci UC San Diego Department of Cognitive Science wrote:

> A common and clever retort, oft heard in this business is as follows:
> Do not confuse simulation with reality. A simulated airplane does not
> actually fly: a simulated intelligence does not actually think. >
> Here is my confusion. There is a fundamental difference between the
> physical operations of an airplane which are simulated in an
> information processing environment and thought processes which are
> simulated in an information processing environment. In the case of
> thought, the actual action and the simulation both involve information.
> The simulation is not that different than the reality, although inside
> a different kind of mechanism.

The notion of "information" is doing a lot of work for you here, and it's in for heavy weather. The formal Shannon/Weaver notion certainly won't bear the weight. But if you extend it further it either becomes so vague that it's no longer clear WHAT the notion is supposed to be picking out, or, worse, it becomes true by definition (and hence vacuously) that what you mean by information processing is simply whatever both the mind and the "information processing environment" can be interpreted as doing.

I'm for a much simpler stance: Interpretations aside, computers just do symbol crunching: They implement a formal program, regardless of whether the program is interpreted as being about an airplane flying, a forest-fire burning or a mind thinking. In none of these cases is the unadorned fact of the matter different: What the computer is really doing is just crunching symbols (i.e., syntactically manipulating symbol tokens on the basis of their shapes, where the symbol tokens can be implemented as scratches on paper or states in a machine) in a way that is systematically interpretable as flying, burning, or thinking.

Now, perhaps thinking is just symbol crunching too. THIS is the logical point where Searle's Chinese Room Argument enters to show that it isn't (and the "Symbol Grounding Problem," Physica D 1990, in press, suggests some reasons why not). That's all there is to it; and hand-waving about "information" is just a symptom of being lost in what I've dubbed the "hermeneutic hall of mirrors," in which one is no longer distinguishing meaningless symbols one is really dealing with from the meanings one has projected onto them. This projection is unwittingly and illicitly parasitic on the meaningful symbols in our head, which could not, on pain of infinite regress, be just an ungrounded symbol cruncher too.

> Harnad and others point out the major problems of grounding one's
> representations before we can hope to simulate human thought. I agree
> completely with this point and have written on the topic myself
> (arguing against the "disembodied mind"), but that doesn't address the
> fact that the simulation of thinking is a different form of simulation
> than the simulation of flying.

One wonders what you could have against "disembodied minds" if you continue to believe that a pure symbol cruncher could think. I wear my anti-disembodiment rationale on my sleeve: Only a grounded symbol system that can pass the Total Turing Test (TTT) can think. The rest are just duds: interpretable as-if thinking, just as all symbol systems and simulations are interpretable as-if SOMETHING (airplanes flying, forest-fires burning, etc.), but interpretable-as-if is not good enough (except for hermeneutics).

For the record, a grounded TTT-scale robot cannot be a pure symbol cruncher; nor is it likely to be just a symbol crunching module connected to peripheral transducer modules. Yet it is this trivial modular caricature that keeps luring the believers in the notion that thinking is just symbol crunching back to the hermeneutic hall of mirrors. For if you (illicitly) equate the brain with a symbol cruncher and the skin and eyeballs with its peripherals, then of course it seems unbelievable that yanking off the peripherals would make the cognitive lights go out. The right way out of this absurdity, however, is not to resurrect the already invalidated symbol-crunching premise, but to abandon the idea that the solution to the symbol grounding problem is just to hook some transducers onto a symbol cruncher. There's much more to grounding than that, as I argue in my papers, and the thinking part of a brain or a grounded TTT-scale robot will NOT be just a disembodied symbol cruncher. It will be hybrid nonsymbolic/symbolic through and through.

> If I simulate a person thinking about a chess game, business decision, or
> mathematics problem, it would be impossible to distinguish the results of the
> simulation from that of a person (a very weak turing test, if you will). And
> sure, because I and others believe that interaction with the world is a
> critical aspect of thought and understanding, let us make sure that the
> simulated mechanism has real sensors and limbs.

Weak Turing Tests are not Turing Tests at all, they are just highly underdetermined toy demo's that mimic an arbitrary fragment of our total performance capacity (and trade heavily on our hermeneutic projections). The number of ways to skin a toy-scale cat is enormous, probably infinite. Not so for a TTT-scale robot, and there the nonsymbolic internal functions that ground any symbolic ones are not just frills: They're INTRINSIC to the grounding, and hence the thinking.

> Notice what I did NOT say:
>
> 1. Simulating thought is not the same as simulating a human. People do other
> things, and the tricky problem of emotions, feeling, and moods remains (these
> seem rather fundamental aspects of humans and one of the distinguishing
> characteristics of our species).

It is possible to adopt a methodological stance that is epiphenomenal, as I have, acknowledging that the only hope for capturing any kind of subjective state -- be it what it’s like to think or what it’s like to grieve -- is to capture everything objective (the TTT) and trust that subjectivity will piggy-back on it. But this is already true of thinking, without worrying about grieving, etc. -- unless you believe in mindless thinking, in which case we’re not talking about the same thing: Thinking in my sense (and everyone else’s, until sci fi fantasies and hermeneutics intervened) requires a thinker, a conscious subject, be he ever so flat of affect...

> 2. This is NOT relevant to the symbolic representational arguments for and
> against connectionism. All existing cognitive mechanisms manipulate
> information, be they connectionist, symbolist, or even quantum mechanist, and
> my point today is only about the information aspect of the process.

And my point is that "information" picks out and explains nothing. As to connectionism, I have suggested a useful natural (but circumscribed) role it might play in grounding categorization in a hybrid nonsymbolic/symboolic system.

> The point which I address now is that a real airplane moves through the
> air, which a simulation can’t do, but a real human having thoughts moves
> information, which a simulation can do.

And if this isn’t hermeneutics then I don’t know what move is.

Stevan Harnad

> From: "Jim Fraser, DTN 381-1552, ZK03-4/S23"
>
> I’m the guy who walked you to the subway after the MIT talks. It was a
> rare pleasure to meet someone whose work I value so highly and discover
> so enjoyable a human being.
>
> I have a little to add by way of feedback to your talks. Though I
> should have predicted it from reading net dialogues on the subject, it
> took sitting in an audience of systems repliers (really, they’re more
> systems pliers) to make clear to me the extent to which these folks
> ignore the difference between simulation and reality. There’s a sort of
> characteristic two-mindedness in which the theoretical model is
> maintained separately from everyday experience. I’d guess it’s
> exacerbated when one spends a substantial fraction of one’s time
> immersed in interaction with computers and/or other rigid rule-based

> mechanisms.
>
> What would have helped the morning talk, I imagined, would have been
> some initial groundwork in the (apparently not) obvious distinction
> between events and their simulations, territories and their maps, with
> inclusion of examples as immediate for the audience as you could
> contrive. There's nothing wrong with the airplane and the forest fire,
> they just came too late, not vividly enough, and almost apologetically.
> It really is hard for me to believe what trouble folks have with this
> distinction - more so in the context of a theoretical presentation,
> where the visceral is about as removed as it ever gets. I imagine
> there's a comfort factor at work somewhere in all this, i.e. that some
> take the distancing of the visceral to be a desirable goal.
>
> I wouldn't make the same comments about the sequence in the paper. I
> think, as you clearly do, that you have a different job to do with a
> live audience. Nevertheless, the talk ran roughly along the same
> sequence as the paper, and perhaps that's not quite right.
>
> In the afternoon talk, I found the groundwork very clearly laid, and
> followed well til the description of the experiments. And it's not that
> I then failed to follow, but that I then began to want to know details
> which would have been inappropriate for the level and duration of the
> talk. In fact, I also failed to capture some of what you did present;
> in trying to recreate your description to a friend, I discovered
> frustrating gaps in my model. Have you written anything on the
> experiments that you'd care to share? I'm interested in details of the
> setup of the nets; my approach to understanding the work would be to
> attempt to replicate it.
>
> You mentioned that you'd send along an e-copy of "Minds, Machines and
> Searle". You mentioned another recent paper of yours in passing, whose
> name I didn't capture; if that exists in e-form and you're willing, I'd
> be happy to have a copy of it as well.
>
> Jim
>

VANISHING INTERSECTIONS AND THE PEEKABOO UNICORN

>
> From: miken@ai.mit.edu (Michael N. Nitabach)
>
> I greatly enjoyed your two talks here at MIT. I'm sorry I wasn't able
> to talk to you afterwards. I must say that the more I think about your
> position versus the pure symbol crunching position, the more I am
> swayed toward the essential role of grounding in producing intelligent,

> understanding, meaningful behavior. With respect to the question I
> asked you about the "problem of the vanishing intersection", I really
> want to believe that this is not a problem. Your point about there not
> being any apparent alternative to grounding in sensory invariants is
> well taken. I think that the question of whether vanishing
> intersections *is* a problem is essential to the success or failure of
> your grounding proposal. As you say there don't seem to be (m)any
> alternatives. I am going to go back to Fodor's Psychosemantics and see
> if I can find anything of substance there. I think you would be
> interested in Dan LLoyd's book Simple Minds, where he develops a
> grounding scheme that is based (as yours is) on detection of sensory
> invariants that can act as determiners of (possibly non-sensory)
> category membership.

As far as I can tell, the only nonmagical alternative to sensory grounding is what I've dubbed the "Big Bang Theory of the Origin of Knowledge" (to which Chomsky and Fodor appear to subscribe). According to this view, the grounding of of our ideas is simply inherent in the structure of matter, including our brains and hence our minds. I find this no better than magic.

> Just a few comments about your connectionist simulation. Although the
> simulations are interesting in their own right, as examples of the what
> a back-prop net does with its "extra degrees of freedom", I am somewhat
> sceptical about its relevance to the symbol grounding problem, vis a
> vis its role as a detector of sensory invariants, and a substrate for
> categorical perception. My justification for this statement comes from
> the "problem of vanishing intersections." If we agree that as the
> "problem" stands or falls, so does your proposal of using *sensory*
> invariants as a base for grounding. i.e. If vanishing intersections *is
> a problem, then the use of sensory invariants as a base for symbol
> grounding is *unlikely* to work. Unfortunately, your simulation
> sidesteps this whole issue; since your network is only sensitive along
> a single dimension of quality, namely length, there isn't even the
> possibility of an intersection vanishing. I would be quite interested
> in seeing the outcome of a similar type of connectionist experiment,
> but where you make the net sensitive to at least two dimensions of
> sensory quality, and thus allow for the possibility of a vanishing
> intersection. Your adherence to a bottom-up view of the construction of
> higher-level categories, however, seems to immunize you from the
> vanishing intersections problem. This is because the only higher-level
> categories that arise in your theory are those that have been
> "discovered" through the operation of your lower level sensory
> invariance detectors. In any event, it would be interesting to see how
> your net architecture performs on a multi-dimensional problem. e.g.
> Could you train a back-prop net to categorize eating utensils into
> knives, forks, and spoons? or something of that ilk. Mike Nitabach

Very good questions, and ones we plan to get to eventually (after all, the whole TTT is the ultimate goal). For now, though, it is not in order to immunize against vanishing intersections that we are concentrating first on the one-dimensional case, but because I believe we have to get a clear understanding of what the net is doing with one dimensional in order to understand what it does with more dimensions.

We will eventually move on to underdetermined multidimensional inputs and categories. But meantime consider this: In order even to DEFINE a category for the net, i.e., in order to have a task in which there is a right or wrong of the matter, and hence something to learn under supervision, the invariants must EXIST in our stimuli, and the feedback must be guided by them. You will reply that this rules out "stimuli" such as "goodness," "truth" and "beauty." My reply is: First things first; just as one dimension must come before many, concrete categories must be grounded before we get to abstract categories and the possibility of discourse. Once discourse is grounded and launched, we can use existing grounded categories not only to define a zebra we've never seen with our senses, but a unicorn that we will never see, and even a "peekaboo-unicorn" (one that vanishes without a trace whenever senses or instruments are trained on it). Now how's that for a perfectly well-grounded category, even though it is not observable with any sense IN PRINCIPLE? Goodness, truth and beauty begin to look a lot less inaccessible as the power of symbol grounding unfolds...

Stevan Harnad

TRIMMING THE FUZZINESS

> From: psych36@ucscd.UCSC.EDU (21045000) Dom Massaro
>
> I just stumbled across the following comments that I wrote when I first
> read your symbol grounding paper months ago: I liked your paper because
> it begins to impose some order on the chaotic controversy between
> symbolists and connectionists. As you might have guessed, however, I
> see the distinction as fuzzy and not clear cut. Turing machine
> equivalence can also be attributed to connectionism, as can systematic
> properties of a formal syntax that is semantically interpretable. For
> example, back propagation is a well-understood formal algorithm. The
> meaning of activations in connectionist systems have to be connected to
> the world in the same way that symbols do. In both paradigms, a leap of
> faith is required to go from stimulus input to processing and from
> processing to behavioral output. These are the psychophysical and
> psychobehavioral correspondences that must be accounted for in addition
> to accounting for intermediate processing between the two.

Only systematically interpretable symbol systems have a symbol grounding problem (it's the problem of grounding the interpretation in what the symbols are interpreted as standing for). If a system cannot even support a systematic semantic interpretation, it has even worse problems (and it is not even a symbol system).

It is not yet clear whether some or all nets are or are not symbol systems. Turing equivalence is a red herring. In one sense of the term, every system is equivalent to some Turing Machine (and most nets are indeed just computer simulations). This can’t be what’s meant. In another sense, a Turing Machine (a computer) can be implemented on connectionist hardware. This can’t be what’s meant either. In a more interesting sense, there is a hypothesis that a net with the right learning rule and training can learn to be a computer. This has not yet been demonstrated, as far as I know.

In my own model, however, nets are only used to do what we aleady know they do well: to find the sensory invariants that allow the net to categorize inputs correctly, according to feedback. The labels of these sensory categories are then the grounded primitives of a grounded symbol system.

Stevan Harnad

FIRST THINGS FIRST

> From: kp@uts.amdahl.com (Ken Presting)
> To: harnad@Princeton.EDU
> Subject: TTT as necessary condition
>
> >SH: But you leave out a crucial distinction that I DO make, over and over:
> >that between ordinary, dumb robots, and those that have the capacity to
> >pass the Total Turing Test [TTT] (i.e., perform and behave in the world
> >for a lifetime indistinguishably from the way we do). Grounding is
> >trivial without TTT-power. And the difference is like night and day.
> ~~~~~~
>
> This sounds like hyperbole. Do you mean to imply that quadriplegia
> entails a loss of mental function? To the extent that sensation and
> action are mental functions, it is clear that diminution of sensorimotor
> capacity is diminution of some aspects of mental capacity. But the
> distinction between (eg) sensation and cognition in general is fairly
> natural (though rough around the edges). Lack of sensorimotor capacity
> entails no lack of memory, no lack of reasoning ability, and (short of
> complete isolation of the brain from the peripheral systems) no lack of
> communicative ability. Briefly, whatever the virtues of the TTT as a
> sufficient condition, it’s inappropriate as a necessary condition.
>
> Ken Presting

A standard mistake. A robot that can do nothing and a blind paraplegic are not the same kind of thing. In particular, the blind paraplegic is not just a symbol cruncher that’s lost its peripherals. Insisting that a robot be able to pass the TTT is the way of ensuring that it’s got the inner wherewithal to qualify for a mind in the first place. THEN worry about blinding and paralyzing it...

Stevan Harnad

THE NEED TO BE EXPLICIT ABOUT SYMBOLS

From: COSCWKY@otago.ac.nz W.K. Yeap.

>SH: It seems to be perfectly legitimate to try to make machines do >intelligent and useful things without caring about whether they are >doing it in a mind-like way. The symbol grounding problem affects such >work only to the extent that grounding may be needed to get the job >done at all.

Yes indeed but at (many) times people get confused as to what their research is about. I have often heard comments from my computer science collegues saying, "everything is AI nowadays..."

>I have the utmost respect for both robotics and early vision (some of >which is symbolic, some not). At the level of generality at which I >was discussing these problems in the paper, the particulars of, say, >Marr's approach were simply not relevant. > >Connectionism was only proposed as one natural candidate for the >learning component of the grounding scheme I described.

I am sure you have but I disagree that you should ignore them. Just as your discussion on connectionism is at a general level without ever mentioning backward propagation or whatever, we don't need any details here either. I think it is particularly crucial here to mention [robotics and early vision] since a sketch of the solution following a "true" AI approach (my definition) would enable (1) a solution to the symbol grounding problem and (2) leave "symbol" undefined IF YOU WANT TO (i.e it can accommodate Pat Hayes's general definition or your more precise definition - in fact it does not really matter how it is defined because THAT IS NOT THE PROBLEM. The problem is how to enable the system itself to have its own semantics.) Here is a sketch of what I call a true AI approach (what Marr called an information processing approach):

It is clear to me that early AI work has demonstrated that symbol manipulation is a powerful mechanism and playing around with the computer to test our ideas is a neat (& fun) thing to do. However, what is missing is that the meanings of these symbols are, as you put it nicely, parasitic on the meanings in our head. We need to know the input, i.e what is happening at the perceptual end, and then ground our symbols to it. Indeed, I believe this is the reason for the shift of AI work towards the "bottom end" in the mid-70s.

Sad to say, few asked the right questions (and many solved engineering problems!) but we do have some excellent work and one example is of course Marr's computational approach (many still confuse his approach with any approach using the computers!). Marr showed the richness of the information computed at the perceptual end, and how next we should ask how we can make sense of it. What constraints are available and how is our world knowledge derived from it?

Example: for vision, how is the notion of object developed from the input? how does one form the concept of a place? for the auditory system, how does the word "ma ma" get identified and related to input from other perceptual systems? Solving these problems essentially amounts to solving the symbol grounding problem. Just one approach (& like connectionism, its success remains to be seen).

If one’s view is that the whole system is a symbol system then the grounding of the symbol is to relate it to information perceived from the environment. If one’s view is that perceptual representations are not symbolic then the grounding of the symbol is based on these representations (essentially what you have suggested). To me, it doesn’t really matter. The important point is that if we proceed along these lines of inquiry, we should be able to ground the symbol somehow. Marr’s work is just one approach and it is only vision (& it has undergone much revision in details already). Connectionism is possibly another (if we can pick out the significant parts from the hype that is going around). But, these approaches provide growing evidence that the mind could be understood (or at least the door to the Chinese room is open)

>SH: The question my paper (and Searle’s) raised was whether standard, >purely symbolic AI -- Strong AI -- could ever succeed in generating a >mind. I gave reasons why it could not.

For those who agree with you (Nitabach, myself, and granting your stricter definition of a symbol system), we do not need to be convinced.

For those who disagree they remain unconvinced (by the discussions).

I thought one interesting aspect of your paper should at least be to provide evidence (in terms of the overall direction) as to how the symbol grounding problem could be solved from work done recently.

Final comment:

If we can ground the symbol, I think the attack of Searle that machines cannot be intelligent is at best premature. I always believe that Searle’s criticism of AI work is unnecessarily harsh - it is a good question but it is not the final word.

[SH: Your ecumenism is commendable, but unless you are prepared to be more specific about what is and is not a symbol and symbol manipulation, it is not clear what you are affirming or denying. After all, any difference can be resolved by simply overlooking it. Symbol grounding is something quite specific, just as symbol tokens themselves are: Symbol tokens are objects (like scratches on paper or physical states of machines) that can be manipulated on the basis of their physical shape in a way that is systematically interpretable as standing for or meaning something. The grounding problem is then simply this: If a particular symbol token, say, "X" stands for a particular object, say, a cat, or a string of symbols, say, "XYZ," stands for a state of affairs, say, the cat being on the mat, what connects the "X" to the cat and the "XYZ" to the cat’s being on the mat (and connects them in such a way as to preserve all the systematic combinatory properties of the symbols and their interpretations), and how does it constrain symbol manipulation over and above the syntactic constraints based on rules that operate exclusively on the shape of the symbol tokens? That is the symbol grounding problem. And although the work of Marr is important and valuable, it can hardly be said to have solved this problem. Stevan Harnad.]

NO HALF WAYS ABOUT IT

From: Drew McDermott SH: What's at issue is like the difference between zero and nonzero: A quantity is not not more or less nonzero, it's either nonzero or it isn't! Both logically and phenomenologically, I think, the kind of continuum you have in mind is simply incoherent. Not so for a continuum between inert and living, whose continuity is in no way problematic, since only objective properties are at issue. But what a nonawareness-to-awareness continuum requires is a graded passage from objectivity to subjectivity, and that's what I claim is incoherent.

What is the intuitive source of this incoherent notion? You already mentioned it: A spurious analogy with degrees of awakeness (sic) in an already known-to-be-aware subject: ourselves.

DM: Did I mention it? Anyway, my main point was not my own unreliable intuitions and analogies, but the methodological point: > DM: But what about the methodological questions? A computationalist could > say, Sooner or later we'll know by inspection of different brains the > degree to which any of them are aware. A "subjectivist" would have to > say, We can *never* really know what it's like to be a bee (say), so we > can never know whether bees are best described as "half-aware," so we > can never know whether there is such a thing as half awareness. I would > have thought you were in this camp, and would remain agnostic on the > question of whether awareness could be partial. SH: I am in the camp that recognizes that the other-minds problem is insoluble: The "methodological" question of whether or not anyone but me is aware is simply unanswerable. Now I've bitten the bullet and accepted that the TTT (possibly augmented to the "TTTT," to include the observable behavior of my neurons) is the most that an empirical theory of mind will ever have by way of objective data, and so I'm prepared to believe that whatever can pass the TTT(T) (i.e., can behave completely indistinguishably from one of us) is aware.

But there's nothing about "degree of awareness" in that. The TTT applies to people. We don't know (and I don't think we'll ever know) the ecology of any other organism completely enough (or intimately enough, for we also lack the introspective empathy to be able to judge, as we can with out fellow humans, what nonhuman mental states might be like) to administer a nonhuman TTT with any confidence, so that's no source of degrees of awareness either. The observed correlates and causal substrates of TTT-passing capacity may turn out to admit of degrees, but nothing at all follows about degrees of awareness from that either. So there you have it.

DM: Exactly my point. All I'm saying is that you must remain agnostic rather than atheistic on the possibility of half-subjectivity.

There are cases one hears about (I think some are in William James) of people who report weird quasisubjective states after brain injuries. That is, they claim (after recovery) to have been aware of the world but unaware that there was any awareness. Or something; it's been a while since I've read any of this. The point is, that it's hard to classify such cases one way or the other. On one hand, the people report being, more or less, thermostats reacting to the world. On the other, they create memories that they later can assign to a coherent self (who was in a weird state at the time). This kind of report is enough to make me very dubious about my own intuitions about subjectivity, and to ask the same humility of others. -- Drew

[SH: Nothing of the sort. This is still subjectivity, nothing "quasi" about it. And awareness-of-awareness is irrelevant. The stubbornly all-or-none thing is whether there is any awareness present at all, not what it is an awareness of, or what it is like. Whatever it is like, it is not the kind of thing that admits of degrees at all. THAT it is like anything at all is the critical property (otherwise known as subjectivity). Either it's there or it isn't. The notion of "half-subjectivity" is just incoherent. Stevan Harnad.]

PLAIN TALK ABOUT ROBOTS

> From: Mike Oaksford mike%epistemi.edinburgh.ac.uk
>
> 1. What have you gained in grounding symbols in features?.
>
> The grounding problem, as I understand your presentation, is that of
> assigning content to the primitive symbols in the symbol system. Thus
> grounding is a new name for naturalising content ascriptions to
> primitive symbolic representations. But features are representations
> also, ie. they are the types or general terms of folk psychology.
> Active congieries of these features represent individuals. Grounding
> one representation in another is circular. In other words,
> Connectionism has no answer to the problem of providing a physical,
> non-semantic reduction of meaning ascriptions, no more than have
> standard symbolic systems.

I'm afraid that in translating it into the terminology of your discipline you have not understood my grounding proposal at all. It's not a matter of mapping one set of representations into another. That's just symbols to symbols, and still ungrounded. The simplest way to see it is as a system that can pick out, identify and describe the objects and states of affairs that its symbols refer to. And the role for connectionism in this system is to learn the invariants in the sensory projection that allow the system to pick out and identify the sensory categories in which all the higher-order categories and discourse are grounded. The system is hybrid, with analog copies of the sensory projections subserving discrimination, invariance-filtered analogs subserving identification, primitive symbols tokens connected to the objects and states of affairs they refer to by the learned invariance detectors, and higher-order symbols and symbol strings composed out of combinations of the lower-order ones, subserving linguistic description.

> 2. Inference is context sensitive, property inheritance does not
> exhaust inference.
>
> It is not simply examples like Tweety etc. which are subject to the
> demand for a context sensitive inferential economy. What of inferences
> such as if the light does not work, the bulb has blown unless the fuze
> has blown, the power is off, the power workers are on strike etc. etc.
> Every rule encoding our common sense knowledge of the world is
> defeasible. No symbolic system can tractably cope with defeasible,
> context sensitive inference, it is an NP-hard problem (cf. McDermott,
> 1987, Computational Intelligence).

No ungrounded symbol system; grounded ones remain to be built and tried. The Total Turing Test (TTT, a robot in the world, indistinguishable form ourselves for a lifetime) is a long way off, but the only valid objective.

> Also observe that on the Holistic view you espouse, your admonishment
> to get literal meaning sorted first is without foundation. On holism
> there is no such thing as literal meaning, since everyone’s inferential
> economy will be slightly different. Hence everyone has different
> concepts not only between individuals but within individuals at
> different times. On this view it is impossible to settle on a fixed
> meaning and thus the notion of literal meaning must be suspect. All we
> can do is transact a shared, public meaning a la Davidson, on which two
> interlocuters agree sufficiently to guide their current course of
> action. Note that this does not mean we have thereby fixed the literal
> meaning of a term (frequently, for the purposes of action, metaphorical
> meaning is not only fixed upon but is preferred, eg. understanding
> electrical current by analogy to the flow of water).
>
> I look forward to hearing from you, Mike Oaksford

I’m afraid you have type-cast me in a philosophical position I neither hold nor even know. I’m just concerned with grounding robots, and taking them increasingly seriously as models of ourselves as they approach the power to pass the TTT. Literal meaning or no, it is the internal wherewithal to get these robots to be able to do what we can do that I think will be the right explanation of how we do it.

Stevan Harnad

[I trust you have received information on how to become a BBS Associate by now.]

MOTOR GROUNDING

> From: Michael Jordan
>
> I enjoyed your recent talk at MIT. I was particularly glad to hear you
> clarify your views on the relationship between learning and
> categorization; I agree that learned categories present the really
> tough problems (although do you really believe that they are the *only*
> ones of interest? Are NP and VP learned?)

In the theory of categorization the learned/learnable problems are the only ones that interest me; no doubt many categories are "prepared" be evolution, but in those cases the interesting part was the evolution. There is no "poverty of the stimulus" agrument for concrete and abstract categories, as far as I know, so there is no need to retreat to a "Big Bang Theory" to explain their origins, and in the case of Universal Grammar.

> Perhaps I am being parochial, but it seems clear to me that motor
> learning problems provide better opportunities for observing the
> phenomena that you are after than do perceptual learning problems.
> In motor systems, the relevant mappings change so thoroughly during
> development (e.g., changes in inertial and other biomechanical

> properties) that it is entirely uncontroversial that the system
> must be highly adaptive. It is also clear that the system must
> form categories in order to make the planning process feasible.
> It must be possible to piece elemental actions into sequences and it
> must be possible to predict the consequences of novel sequences.
> (Take the example of current work in phonology; I think that it is no
> accident that they are now making progress again on the notions of
> ‘‘feature’’ and ‘‘phoneme’’ by focusing so heavily on articulatory
> representations). Michael Jordan

I completely agree that motor learning is very important. But remember that learning to assign an arbitrary identifying label to a class of inputs IS a special case of motor learning! Moreover, sensorimotor interactions with the world probably provide more of the interesting "preparation" for categorization than even evolution does. Nevertheless, there is a big difference between analog and arbitrary motor responses, as I discussed in my second talk, and at the heart of the categorization problem is arbitrary labeling, which -- apart from any preparation there may be from evolution or analog activity in the world -- is based on sensory (or, rather, sensorimotor -- as the person who asked the Gibsonian question reminded me) invariants.

Stevan Harnad

CONVERGING ON CONCSIOUNESS THROUGH PERFORMANCE CAPACITY

> From: kp@uts.amdahl.com (Ken Presting)
> Subject: TTT as a necessary condition, only for robots
>
> I wrote:
> >> . . . Do you mean to imply that quadriplegia
> >> entails a loss of mental function?
> >> . . . Briefly, whatever the virtues of the TTT as a
> >> sufficient condition, it’s inappropriate as a necessary condition.
>
> You replied:
> >A standard mistake. A robot that can do nothing and a blind paraplegic
> >are not the same kind of thing. In particular, the blind paraplegic is
> >not just a symbol cruncher that’s lost its peripherals. Insisting that
> >a robot be able to pass the TTT is the way of ensuring that it’s got
> >the inner wherewithal to qualify for a mind in the first place. THEN
> >worry about blinding and paralyzing it...
>
> If I follow you, (which I’m not sure I do) you are proposing that we must
> wait for robotic technology to advance to the point where we can build
> a full scale TTT-capable device before we can confidently attribute mental
> properties to simpler variants with fewer sensorimotor functions. To
> check my understanding, let me speculate on your reasons for rejecting
> blind paraplegics as counterexamples to the necessity of the TTT. Is it
> that we have live full scale TTT-passing examples of human beings that

> gives us the methodological luxury of applying only the TT to our
> handicapped fellows? I would appreciate some elaboration on this topic.

That's just about right. Let me put it more plainly: We know human beings have minds (though we don't know how they work), so it's a pretty safe bet that human beings without eyes or muscular control still have minds (even though we still don't know how their minds work), because it's a pretty safe bet that their minds are not in their eyes of their muscles but in their brains -- and, to repeat, we don't know how those work (although, thanks to Searle and others, we do know that, however they may work, it can't be just by crunching symbols).

I am NOT saying that if we get a TTT-scale robot then "we can confidently attribute mental properties to simpler variants with fewer sensorimotor functions" -- except in the trivial case where we know we are generating the variants merely by lopping off the eyes or muscle control* (and there may well turn out to be ways to be confident that that's what we're doing). But other "simpler variants" -- like robots that can ONLY play chess, or ONLY prove theorems, or ONLY manipulate blocks, etc, are just back to the old toy-domain, in which there is no mind at all. (The underlying "convergence" argument -- from "Minds, Machines and Searle" -- is simply that the degrees of freedom for generating toy fragments of our performance capacity are much larger than the degress of freedom for generating our total performance capacity.) Even TTT's for nonhuman species turn out to be a methodological problem, largely because we lack the ecological knowledge to ascertain just what their Total robotic capacity is, and when they've achieved it, and partly because we lack the innate empathy (knowing what it's like on the inside) that makes the TTT so sensitive with our own kind.

Passing the human TTT is neither demonstrably a necessary condition (animals surely have minds, and they can't pass the human TTT, for example) nor is it a sufficient condition (the TTT robot may NOT have a mind). It is merely a methodological criterion for mind-modeling -- and, I claim, the best one we can adopt. (The TTTT -- our total bodily AND neural performance capacity -- which amounts to as much as empiricism can ever offer, looks like a stronger Test. I argue, however, that the TTTT is (1) stronger than we need, because the remaining degrees of freedom among equipotential TTT systems are just implementational, (2) only the TTT, i.e., performance capacity, can pick out which aspects of neural function are RELEVANT to having a mind and (3) the TTTT is no guarantor either, because of the other-minds problem.

> The reason I'm harping on this topic is that your proposed solution to
> the symbol grounding problem seems to operate cumulatively, which is a
> mismatch (to my mind) with your all-or-nothing stance on the TTT.
>
> Ken Presting

I don't think that TTT-passing power is just the sum of all the toy capacities we have. I would say grounding is convergent rather than cumulative, at some critical point (demarcated by the human TTT, rather than animal TTTs, for the methodological reasons I mentioned) calling for internal causal substrates that evolution, at least, does not seem to have managed to muster without making them conscious. In other words, I doubt that we are any more likely to be able to stumble onto a successful mindless variant with the same performance capacity than the blind watchmaker was -- which squares with my argument that the degrees of freedom for implementing such powers must be small.

Stevan Harnad

* This is just hypothetical talk. I think it would be as wrong to abuse TTT robots as it is to abuse animals.

THE ARBITRARINESS OF QUANTITATIVE THRESHOLDS

From: harnad (Stevan Harnad) To: block@cogito.mit.edu Ned Block

The position you took seemed ambiguous to me. You did and didn't want to agree that your rejection of Searle's conclusion was based merely on the hypothesis that at some point (in the speed, size, or complexity) of symbol manipulation some sort of "phase transition" into mind-space might occur. I see that as pure sci fi on all currently available evidence.

Your other point was again a somewhat ambiguous endorsement of the analogous claim that Searle's Argument is moot, because Searle couldn't manipulate the symbols fast enough, but that if he COULD do so, then maybe he WOULD be able to understand Chinese. I find this equally fanciful on all available evidence.

All of this, I'm afraid, falls under the "heremeneutical hall of mirrors" diagnosis, according to which the sheer INTERPRETABILITY of it all is so compelling that the symbol-crunching thesis keeps getting accorded the benefit of the doubt, at a mounting counterfactual price. This may be the lure of the TT itself (which is why Searle rejects the TT), and perhaps even of the TTT (which I endorse), but the latter at least has the virtue of not being vulnerable to Searle's Argument -- and of being just one short of the TTTT, which is the bottom line not only on the other-minds problem, but on empiricism and objectivity itself. The hermeneutic buck stops there too.

Cheers, Stevan

SYMBOL SAVING

> From: djoslin@bbn.com David Joslin:
>
> SH> A standard mistake. A robot that can do nothing and a blind paraplegic
> SH> are not the same kind of thing. In particular, the blind paraplegic is
> SH> not just a symbol cruncher that's lost its peripherals. Insisting that
> SH> a robot be able to pass the TTT is the way of ensuring that it's got
> SH> the inner wherewithal to qualify for a mind in the first place. THEN
> SH> worry about blinding and paralyzing it...
>
> Not so fast. Let me try to summarize the discussion we had after your
> first talk at MIT. Tell me if I'm misrepresenting you.
>
> 1. You claimed that Searle's Chinese Room argument shows that a
> symbol cruncher could not genuinely understand language. I still
> have some questions about this, but let's grant it for now.

>
> 2. You claimed that a robot that passed the TTT would genuinely
> understand some language.
>
> 3. You agreed that if a symbol cruncher plus some simple transducers
> could pass the TTT, then it would, based on (2), genuinely understand
> some language. You expressed doubts about whether this was possible,
> but let’s assume for the moment that it is possible. (By "simple
> transducer" I mean one that produces a symbolic output, such as a video
> camera with a digital output, or a microphone connected to an A/D
> converter. The point is to make as much of the process symbolic as is
> physically possible.)
>
> 4. I asked what would happen, given (3), if we turn off the camera
> and other transducers. You replied that, since the one thing we
> are certain of is (1), and since turning off the transducers leaves us
> with just a symbol cruncher, turning off the camera would, in
> fact, cause the robot to cease to understand language!
>
> Dan Dennett once said "The problem with a _reductio_ argument is
> that sometimes your opponent will decide to embrace the _absurdum_."
> The idea that turning off the transducers leads to the loss of
> understanding strikes me as utterly absurd, and a good reason to
> give up (1) or (2), or possibly to deny that a symbol cruncher plus
> simple transducers could pass the TTT. This last one would
> sound suspiciously like a theory-rescuing maneuver, but it would
> still seem preferable to the conclusion you suggested at MIT.
>
> What are your thoughts on this now? David Joslin

No theory saving at all: It seems very unlikely that just cutting off the transducers should turn off the mental lights, but the hypothesis was counterfactual in any case, so let the conclusion stand. (We know little enough about what will and won’t generate a mind to be able to live with a lot of unlikely sounding possibilities; Strong AI itself was such a one, even before it met Searle and the Chinese Room.)

A more likely possibility, however, is that what generated the unlikely conclusion was a false premise. My candidate for the locus of the falsity is the premise that the TTT could be successfully passed by just a symbol cruncher hooked to peripherals. I am not resorting to sci fi or theory saving here, for there are perfectly reasonable alternative candidates for the kind of system that might indeed pass the TTT (among them my hybrid symbolic/nonsymbolic model), and they do not lead to the same unlikely conclusion.

By contrast, the "Systems Reply" is indeed a case of theory saving at any counterfactual price. Again, the premise is counterfactual: that a symbol cruncher could successfully pass the TT. I find this unlikely, but let’s see where it would get us if it were true: To the Chinese Room, and the very unlikely conclusion that Searle would contract multiple personality disorder from just memorizing a bunch of symbols. Unlike in the case of mind modeling, however, where we have no evidence one way or the other about what kind of process does or does not give rise to a mind, here we have the empirical evidence (test it if you like) that memorizing symbols is NOT the kind of thing that causes multiple personality. So we have two choices: We reject the conclusion that Searle would have two minds (i.e., we reject Strong AI) or we reject the premise that symbol crunching alone could pass the TT in the first place (but this is likewise to reject Strong AI).

So where's the theory-saving?

Stevan Harnad

ON SUBHUMAN CONSCIOUSNESS AND THE SUBHUMAN TTT

> From: Jeff Inman
>
> (1) The TTT:
>
> I think I understand your reasons for devising the TTT, namely that you
> want it made clear that total sensory involvement is necessary, if we
> are to be able to appreciate the symbolic structure of a robot's
> functioning. However, isn't it the case that no robot could ever really
> pass this test because everything that passes the test is, by
> definition, a person? Presumably, you would argue that the robot's
> constructor could know that one of the "people" in the room is a robot,
> and that if the "real" people all tried and failed to guess which one,
> then the test passed. Now, let us suppose that the constructor forgot
> which one the robot is. Can she figure it out? In other words, does the
> robot's constructor know of any special weaknesses that would allow her
> to fail the robot in the TTT? If not, then mustn't we argue that what
> we have is a room full of people?
>
> Personally, I don't agree with my assertion here, but I'd be curious to
> hear your response.

You should have agreed with your assertion, because it's quite right. To accept the TTT as the best (and only) objective indicator we can have for anyone else's having a mind is to accept that any candidate that passes it is a person. Period. To reject it is to imply that we know better, or have some other means of judging such things. In reality (and that includes everyday life) we do not.

If you have followed this, then you must see why you have to stop thinking of a "robot" as something that doesn't have a mind (or is not a person). According to the TTT criterion, some kinds of robots must have minds, because we are such robots. A robot is just a causal physical system with certain performance capacities. Its being man-made or its internal functions being completely understood (say, deterministically) are no grounds for denying that it has a mind (though it may not have one, of course, because the TTT is no guarantor -- nothing is).

And the TTT is as conclusive as it gets, not because of "total sensory involvement," but because it's as much as we can ask for and as much as we ever get by way of objective evidence of anyone's having a mind.

[Note: The "TTTT" -- which extends Turing indistinguishability not only to the body, but also to all parts of the body, including neurons and biomolecules, is a stronger test (in fact, it is simply equivalent to empiricism), but certainly not the one we ordinarily use with one another. In fact, it's probably TOO strong, methodologically speaking, ruling out systems with minds that simply differ implementationally, in ways that do not have anything to do with having or not having a mind. In practice, however, I conjecture that the difference between the TTT and the TTTT will not turn out to be important, because all the hard problems will already be solved by the time we can get ANY system with the performance capacity to pass the TTT; the rest will just be the fine tuning.]

> (2) Continuum of awareness:
>
> I've yet to run into anyone who finds merit in Julian Jaynes' theses
> [anyone besides myself, for despite my basic disagreements with Jaynes,
> I still admire the scope and basic tenets of his idea) and I myself
> find him to be a reductionist (contrary to myself). I found the idea
> fascinating that (what Jaynes identifies as) "consciousnes" --
> apparently, the existance of a subjective, internalized space within
> which one has a notion of one's self as an entity in the world -- could
> be argued to have erupted in the heads of humans only 2500 years ago.
> You could translate this perhaps as saying that humans had achieved
> sufficiently complex grounding structures that they could begin
> grounding the concept of 'I', a self. Let me state again, I have
> problems with some of Jaynes ideas. For example, it seems to claim that
> humans have somehow more "consciousness" than other animals, and I
> don't believe that; rational symbol-manipulative power, maybe.
>
> The point that I thought might involve this discussion is that if you
> claim that humans are "aware", we might ask "since when?". If we are
> positing evolutionary forces, we might suppose that there were earlier
> forms of "awareness" than that which we know now. Drew McDermott's
> ideas about bee awareness are well taken.

Unfortunately, there's a contradiction or an incoherence above. I'm perfectly prepared to believe (in fact I do believe) that bees and other nonhuman species are conscious, in the sense that they have subjective experiences, whatever their specific quality might be. That's what having a mind means. But then this means that the really radical differences are much further upstream than Julian Jaynes (a good friend, whom I admire enormously, by the way, despite disagreeing vehemently with him) places them. He and others are speaking about certain KINDS of subjective states and capacities -- those involving reflection, usually mediated verbally, or what the philosophers called "awareness of awareness." But this 2nd order awareness is a mere frill to me, compared with the question of the origin and basis of awareness itself: 1st order awareness, the capacity to have any kind of subjective experience at all.

And I've argued strongly that it's useless to look for an evolutionary explanation for consciousness -- in terms of its adaptive value, or what have you -- for the simple but insuperable reason (follow carefully) that EVOLUTION IS JUST AS BLIND TO ANY POSSIBLE DIFFERENCE BETWEEN A CONSCIOUS ORGANISM AND ITS TTT-DOPPELGANGER! They're functionally equivalent (TTT-indistinguishable) in every respect. So the only "Darwinian" factor is whatever the internal functional wherewithal to pass the TTT is. [This is a subtle point, and takes some thought, but it's a solid one. It’s one of the reasons I’m a "methodological epiphenomenalist," as explained in a 1982 paper called "Consciousness: An Afterthought.]

> I don’t see how you can claim
> that the other minds problem is insolubale and yet also say this:
>
> SH: But there’s nothing about "degree of awareness" in that. The TTT
> applies to people. We don’t know (and I don’t think we’ll ever know)
> the ecology of any other organism completely enough (or intimately
> enough, for we also lack the introspective empathy to be able to judge,
> as we can with out fellow humans, what nonhuman mental states might be
> like) to administer a nonhuman TTT with any confidence, so that’s no
> source of degrees of awareness either.
>
> I think you’re cheating by claiming to be able to empathize with other
> humans, yet refusing to empathize with other species, mechanisms, etc.
> We all know how easy it is to begin to empathize even with the simple
> robots of today, which seem to "want" to do X, or be "too stupid" to
> figure out Y, etc.

I do empathize with other species. I believe they have minds just as fervently as I believe other people do, and I am accordingly a vegetarian. But empathy is only one side of the TTT coin (and it does diminish as one moves to invertebrates, unicellular organisms and biomolecules); the other side is methodological: We simply do not know the full performance capacity of any creature (even simple creatures) nearly as fully and intimately as we know our own; indeed, our cross-species empathy is based entirely on the degree of similarity that there is between our own performance capacity (including dynamics of "appearance," such as facial expression, movement and vocalization) and that of other creatures (and how it maps onto corresponding inner states that we each know at first hand); and I don’t think we ever will know.

This doesn’t mean that subhuman TTT-models can’t be fully validated. I think they will be necessary early steps and approximations along the road to the full human TTT. But we cannot hope that they will ever have the methodological force that the human TTT has. I mean, we’ll never know enough ecology to be as sure that a robot-axolotl has passed the axolotl TTT -- despite the fact that its performance capacity seems so much simpler than ours -- as we will be when that robot spends a lifetime indistinguishably in our midst.

And I certainly can say the other-minds problem is insoluble, for irrespective of what "definitions" or "criteria" I adopt, the only way to be sure anything has a mind is to be that thing, and I have that privilege only in one single case: my own. All others are separated by an impenetrable barrier. That’s the insoluble other-minds problem.

> (3) Chinese rooms
>
> [Maybe you’ve been through this one before. I hear that the morning
> lecture at MIT addressed some of this.]
>
> Why should we assume that Searle knows anything about whether the
> Chinese Room understands chinese? Searle is merely an active component

> in the chinese room's operation, but not necessarily the component that
> must have awareness of its understanding. The person to ask is the
> chinese room itself. Presumably, it would answer "yes, of course" .. in
> chinese.

Ah me, the "systems reply" again, in its standard incarnation. Yes I gave the answer in the morning lecture; so did Searle, in his paper. But one has to be prepared to draw the logical consequences for one's beliefs, and the systems-repliers cling too strongly to their original premise to let go of it in the face of counterevidence. They just keep resurrecting it with more and more counterfactual science fiction whenever it is cast in doubt. Here we go again:

The right version of Searle's Argument to focus on is the one where he has memorized all the symbols and all the rules for manipulating them. Then it's clear that there's no one to point to but Searle himself (so forget about the mind of the room). Then the only recourse for the systems-replier is that there must therefore be two minds in Searle, dissociated from one another: an English one, and a Chinese one (that's the second "system"). The simple answer to this is that although there is clinical evidence for the existence of multiple personality disorder -- two or more minds co-existing in the same brain, with one not knowing about the other -- there is unfortunately no evidence whatsoever (nor is there ever likely to be) that one of the possible etiologies for this disorder is memorizing a bunch of symbols. Early sexual abuse maybe, but not symbols. To believe otherwise is pure sci fi.

Stevan Harnad

To: Recipients of the Symbol Grounding Discussion list.

After a lull, the List is active again. Please let me know if you want your name removed from the list. SH.

(1) UNDERSTANDING HOW TO MANIPULATE SYMBOLS (Presting) (2) THE LOGIC OF COUNTERFACTUAL CONDITIONALS (Joslin) (3) THE UNHELPFULNESS OF NEUROSCIENCE (Nitabach) (4) PROOF BY ASSUMPTION (McDermott) (5) DON'T BLAME DESCARTES (Preston)

ON UNDERSTANDING HOW TO MANIPULATE SYMBOLS VS UNDERSTANDING WHAT THE SYMBOLS MEAN

> From: kp@uts.amdahl.com (Ken Presting)
> Subject: Symbol Grounding Discussion
>
> You must be joking with the "Multiple Personality" comment on the
> System Reply. No clinician would be tempted even for a moment to
> diagnose MPD in the case where Searle memorizes the rules.

Of course no clinician would; it's the Systems-Replier who thinks one would (or should). Searle's explanation of what actually happened ("I just memorized a bunch of meaningless symbols...") is perfectly plausible, and, what's more, true!

> Since the Chinese Room has come up yet again, let me pass along my
> account, which I think is directly relevant to the symbol grounding
> problem.
>
> Note first that the symbolically encoded rules given to Searle are
> identical to the program executed by the computer (if they aren't,
> they might as well be). Now, Searle reads the rules, understands
> them, and intentionally obeys them. Therefore the rules are *not*
> devoid of semantic content, and Searle's Axiom 1 (Sci.Am. version)
> is false.

No. The rule says, if you see a squiggle, give back a squoggle. The foregoing sentence has semantic content all right, and Searle would understand it, but it's not THAT semantic content that's at issue. In the Chinese case, it's the meaning of the Chinese symbols (because that's what Strong AI-ers are claiming that the system which is implementing the program is really understanding) that's at issue, but the same problem would arise if the computer (or Searle) were doing arithmetic or playing chess. If Searle were given 2 + 2 = ? in, say, octal code, he'd be doing squiggle squoggle -- in fact, he might be reading exactly the same rule as at the beginning of this sentence, but it's meaning (which he would not know, of course) would be different. THAT's the semantics which is at issue; not the semantics of the squiggle-table (which, by the way, Searle both understands AND implements, whereas the computer only implements).

> Grounding comes in when the rules are loaded into a real computer,
> and the question of the semantic content of the rules arises again.
> Now the rules, as they exist inside the machine, determine the
> causal process of the machine's operation. The correspondence
> between the syntax of the rules and the operation of the machine is
> precise (for obvious reasons). Therefore the symbols used in the
> rules *are* grounded, and this holds for every program.

I'm afraid you've misunderstood the grounding problem. To implement a program (which is all you're talking about here) is decidedly NOT to ground its symbols. If it were, there would be no grounding problem at all! No, the symbols are also presumably semntically interpretable as meaning something more than "If squiggle then squoggle." Whatever that something is (Chinese statements, arithmetic statements, chess moves, scene descriptions), THAT's what's not grounded. For example, there may be a line like "The cat is on the mat." The meaning of that symbol string, and its components, is ungrounded. It gets its meaning only from our interpretations.

> A few disclaimers. I'm not arguing (here) that either the CR or
> the computer understand Chinese, I'm only arguing that Searle's
> argument fails. Also, I'm making no claims about the groundedness
> of any Chinese symbols as used by the CR or the machine, the only
> groundedness claim is for the symbols used to state the rules. I'd
> like to take up these issues later, however.
>
> Ken Presting

But no one was particularly concerned with the statement of the syntatic rules. If Searle didn't understand those then the only way he could implement the program would be to have someone or something else PUSH him through the motions! From Strong AI's standpoint, it's enough that he is actually going through the symbol manipulation seqeunce, i.e., he is indeed an implementation of the program that putatively understands Chinese. If you're not interested in whether or not he understands Chinese then you've changed the subject.

Stevan Harnad

ON THE LOGIC OF COUNTERFACTUAL CONDITIONALS

> From: djoslin@BBN.COM (David Joslin)
> Subject: Turning off the lights
>
> In a nutshell, the four points in the argument from my previous message were:
>
> 1. The CR argument shows that a symbol cruncher could not
> understand language. (Assumed for the sake of argument.)
> 2. A robot that passes the TTT would genuinely understand
> language. (Your claim.)
> 3. If a symbol cruncher plus simple transducers passed the TTT,
> it would understand language. (from 2)
> 4. Turning off the transducers in (3) would cause the robot
> to stop understanding language. (from 1)
>
> I argued that (4) was so counterintuitive that either you needed to show
> that (3) was impossible (i.e., that a symbol cruncher plus simple transducers
> could not pass the TTT), or else give up (1) or (2).

I don't see why. It seems much more sensible to take 1, 2 (the TTT being just a strengthened form of the TT, immune to 1) and 4 together to imply that the "If" clause of 3 probably cannot be satisfied by a real system. Besides, supposing that turning off the transducers turns off the mental lights still sounds more plausible than that Searle understands Chinese without knowing it. But really, it's not a good idea to walk out too far on limbs constructed of counterfactual conditional upon counterfactual conditional. And it certainlly isn't the locus from which to try to reconstruct the tree...

> I want to come back to the idea that attempting to hold onto (4),
> even though we both agree that it "seems very unlikely," is a
> theory-saving maneuver, but I'd first like to get some clarification
> on your objection to (3).

Note in passing that the only "theory" at issue is Strong AI, i.e., the theory that thinking is just symbol crunching. I described some of the sci fi that people resorted to in order to try to rescue that theory from Searle's argument as theory saving. I can't be theory saving because I don't have one. I just want to reject what's demonstrably absurd and not take leave of common sense unless it's absolutely necessary.

>
>sh>A more likely possibility, however, is that what generated the unlikely
>
>sh>conclusion was a false premise. My candidate for the locus of the
>
>sh>falsity is the premise that the TTT could be successfully passed by
>
>sh>just a symbol cruncher hooked to peripherals. I am not resorting to
>
>sh>sci fi or theory saving here, for there are perfectly reasonable
>
>sh>alternative candidates for the kind of system that might indeed pass the
>
>sh>TTT (among them my hybrid symbolic/nonsymbolic model), and they do not
>
>sh>lead to the same unlikely conclusion.
>
> Okay, let's imagine a hybrid symbolic/nonsymbolic robot that passes the
> TTT. Let's assume that it is an entirely electronic robot, except, of
> course, that there have to be transducers to convert electronic signals
> to movements, to convert photons to electronic signals, and so on.
> Pick any electronic sub-circuit in this robot's "brain." What prevents
> me, in principle, from replacing that circuit with a digital circuit
> plus A/D and D/A converters, that duplicates the behavior of the
> original circuit to whatever precision is required? Unless, somehow,
> infinite precision is required in some signal level, I don't see any
> problem, given sufficiently fast components and enough bits in the
> digital representation. So we repeat this process over and over. If we
> have two digital circuits connected by D/A and A/D pairs, we simply get
> rid of the converters and use a digital bus. And so on. Repeating the
> process, we eventually end up with a digital machine connected to
> simple transducers -- exactly what was required in step (3) above. And
> since the original robot passed the TTT, if the new circuits are true
> functional equivalents then the new robot will also pass the TTT.
>
> Which brings us right back to the situation above. So while a hybrid
> robot may not lead us *directly* to the unlikely conclusion, it
> does seem to get us there indirectly.
>
> You've said recently that a TTT-scale robot will not be just a symbol
> cruncher plus simple transducers, it will be "hybrid nonsymbolic/
> symbolic through and through." I don't understand what you mean by this.
> The above sketches out a reason for thinking that the existence of a
> hybrid (electronic) robot that passes the TTT implies that, given
> sufficiently fast components, a symbolic robot (plus transducers) that
> passes the TTT must also be possible. (If you want to object to my
> assumption of electronics, we can address that issue.)
>

> (I want to make sure that we don’t mean different things by
> "transducers." I think of the transducer as the part of the robot
> that, for example, converts photons into whatever kind of signal is
> used internally. For an electronic robot, the visual transducer might
> be a CCD array. In a human, the visual transducers are the rods and
> cones that convert photons into nerve impulses or whatever. At times,
> in trying to understand the way you use the word "transducer," you seem
> to be including much more circuitry in the transducer. True? If so,
> how much, and why?)
>
> David

I like your reductio, so I will try to put exact bounds on what it does and doesn’t show:

(a) First, let’s assume we mean the same thing by transduction. My claim is that a lot more of the work has to be done directly in analog form, without going to A/D, symbol crunching, and then back to D/A. Now, I have often rejected empty hand-waving on the part of others about speed, capacity and complexity when they have tried to use them to support some sci fi fantasy they’ve conjured up to protect a pure symbol cruncher from Searle’s arguments or my own. So I must be careful not to hand-wave about speed, capacity or scale myself in the same way. The illicit way is simply to conjecture some sort of an arbitrary phase transition (in speed, capacity, or complexity) that breaks out of the physical into the mental, or vice versa. This I certainly won’t do. I suggest, though, that there are perfectly natural and plausible reasons for believing that some kinds of processes become prohibitively slow, big or complex if done symbolically rather than in analog. There may be purely physical reasons why, if every process is cashed into symbols except for its outermost interface with the world, some kinds of processes may become intractable, and organisms and their brains may be made up of a lot of those. You’re asking for something analogous to replacing every TTT-relevant biological process in the brain and the body by a symbolic simulation, cashed in only with simple transducers at the periphery, so as to still work, TTT-style, in the real world. Well, I can’t prove it, but I sure doubt it can be done.

(b) Perhaps a few examples of the analogous kinds of systems on which my doubts are based will help: My standard airplane and furnace examples will do. Twist it any which way, airplanes and furnaces are not nontrivially replacable by symbol crunchers so as to still yield devices that fly and heat (i.e., pass the plane and furnace "TTT"), and whose critical properties reside in the symbol crunching rather than the transduction. Well, I think the same is true of perceptuomotor function, on which most of the TTT is based, and in which symbols must be grounded, according to my own iconic/categorical candidate for solving the symbol grounding problem (but note that I’m not particularly trying to "save" my own theory in any of this argument). It makes about as much sense to try to transform my grounded system into just trivial peripherals plus symbol crunchers as it does to do so with planes and furnaces.

(c) One of the reasons it would be difficult to do the symbolic reduction you envision is related to the TTT itself: A successful TTT-passing robot has to have the capacity to perform indistinguishably from a person for every possible input contingency. To capture all of those symbolically, it is not enough to symbolify all the innards of the robot except some thin, trivial peripherals, it’s necessary to symbolify the world too, and everything it might do to those peripherals. A tall order, it seems to me, for symbols.

(d) But if the likely size, speed and complexity of the requisite symbol cruncher still doesn't seem counterintuitive to you at this point, then I don't see why I should see it as any more counterintuitive than proposition (4), which you feel should make me abandon the rest of my premises.

(e) Finally, I don't think your summary of the Argument that you gave in (1) - (4) gets all the conditionals in place. Here's how the Argument really goes:

(i) IF a pure symbol cruncher could pass the TT, THEN Searle could implement it too; therefore, neither of them would understand.

(ii) IF a robot could pass the TTT, THEN there would be no basis for denying that it could understand.

(iii) IF a symbol cruncher plus simple transducers could pass the TTT, THEN it would understand.

(iv) Turning off the transducers would cause (iii) to stop understanding.

Note all the conditionals, each of which could, in principle, be counterfactual, i.e., as unrealizable as "IF we could trisect and angle...". I happen to doubt the IF in (i), because of the symbol grounding problem. My confidence in the doubt is strengthened by the reductio provided by Searle's thought experiment. So confidence in pure symbol crunchers has gone way down at this point in the argument.

Next we have a more plausible conjecture, (ii) that SOME kind of system (and necessarily not a pure symbol cruncher) might pass the TTT. Fine; I see no reason to deny this. We, for example, seem to fill the bill, no strings attached. So there certainly seems to be no principled reason for saying a priori that such a system would not understand. This is not to say that a reason might not arise someday, as in the case of symbol crunching and Searle's argument. But right now there's no reason in sight, and I'm willing to accept the consequences: TTT-indistinguishable systems all understand.

Now we have another conjecture: IF the symbol cruncher, discredited in (i), could pass the TTT merely by tacking on some trivial transducers, THEN it would understand. It seems to me that there are several ways to go on this, one being to doubt that such a system could pass the TTT in the first place (as I do, because of the Symbol Grounding Problem); but if, as in (i), we accept the premise arguendo, then the consequence seems tolerable: IF this were possible, THEN it would understand.

And (iii)'s corollary (iv) seems tolerable too (or at least no more intolerable than (iii)'s antecedent IF clause itself): IF such a system understood, THEN it would stop understanding if deprived of its peripherals, because of (i). I see nothing wrong with that as a conclusion from a probably counterfactual premise. But common sense suggests that the truth is probably that it follows from (i) that (iii)'s IF clause cannot be satisfied, hence (iv) need not be considered. It certainly does NOT imply that the (i) is false; and I find that it has very little bearing on (ii), i.e., the TTT.

Searle successfully discredited the TT in the special case of a pure symbol cruncher because of the privileged periscope it gives you on the usually impenetrable "other minds" problem: Since EVERY implementation of a pure symbol system S that allegedly has property P must have property P (that's just "symbolic functionlism"), Searle can implement the system and confirm that understanding is not one of its properties after all. The TTT is not vulnerable to any such demonstration, so it’s as sound as it ever was. (I of course don’t claim it is infallible or necessarily true; it’s simply a stronger methodological conjecture than the discredited TT.)

Now I admit that there’s an uncomfortably high level of counterfactual conditionality in all this, but I think we can still keep our bearings: Pure symbol crunching and the TT are safely discredited (unless you want to believe in multiple minds, induced by memorizing meaningless symbols). The TTT is safe for now, and a pure symbol cruncher plus simple peripherals seem unlikely to ever be able to pass it, probably for capacity reasons. Complex transducers (probably consisting of dynamical systems) plus not-so-complex grounded symbol crunchers are more likely candidates for successfully passing the TTT. Remove those transducers any symbol cruncher that remains will not have a mind.

Stevan Harnad

ON THE UNHELPFULNESS OF NEUROSCIENCE IN EXPLAINING OUR PERFORMANCE CAPACITIES

> From: miken@ai.mit.edu (Michael N. Nitabach)
>
> In your recent reply to Jeff Inman:
> SH>[Note: The "TTTT" -- which extends Turing indistinguishability not only
> SH>to the body, but also to all parts of the body, including neurons and
> SH>biomolecules, is a stronger test (in fact, it is simply equivalent to
> SH>empiricism), but certainly not the one we ordinarily use with one
> SH>another. In fact, it’s probably TOO strong, methodologically speaking,
> SH>ruling out systems with minds that simply differ implementationally, in
> SH>ways that do not have anything to do with having or not having a mind.
> SH>In practice, however, I conjecture that the difference between the TTT
> SH>and the TTTT will not turn out to be important, because all the hard
> SH>problems will already be solved by the time we can get ANY system with
> SH>the performance capacity to pass the TTT; the rest will just be the
> SH>fine tuning.]
>
> This conclusion provides a "new, improved" justification for the centrality
> of AI as a sub-discipline in cognitive science. That is, all the hard
> problems will be addressed in achieving the TTT (i.e. AI). Specifically, you
> have discarded purely symbolic functionalism and adopted a stance of robotic
> functionalism. Even from this new stance, however, the view of nervous
> system physiology and morphology as an implementational detail, independent
> of the functional level, persists. What is your evidence for this view?

Only the fact that nothing that WORKS, i.e., that actually generates successful (nontrivial) performance capacity, has ever come out of neuroscience. Nor is it likely to, because I think the kind of peek and poke possibilities (stimulation, recording, ablation, pharmacology, staining) that are available will not reveal the functional basis for the brain’s performance capacities (TTT). It’s as if a paleolithic culture were given a fleet of jet planes and told how to fly and even how to repair them, but not how they work (i.e., not the laws of aerodynamics and the principles of aerospace engineering). Now, which do you think would be the more likely path to their arriving at an understanding of how airplanes work: By peeking and poking at them, the way neuroscientists do? Or by recapitulating western science and engineering, discovering the laws of physics and figuring out (by trial and error) the principles of aeronautical engineering? I'd bet on the latter.

By the way, if by AI you mean standard AI (symbol crunching), I certainly don't think I've provided a justification for it. If by "AI" you mean any which kind of robotic modeling, then I'm inclined to agree; and I think computer simulations of flight might help our paleolithic proto-engineers too (if they were given and shown how to use computers as miraculaously as they were given jets).

Stevan Harnad

ON PROOF BY ASSUMPTION

> From: Drew McDermott
>
> I just want to reiterate a point about logic that I've made before.
> I am sure that if I could just get you to see this one point, your
> certainty about the Chinese Room would waver. Of course, that's a
> big if.
>
> First, the obligatory quote:
>
> > From: Jeff Inman
> > ...
> > (3) Chinese rooms
> > ...
> > Why should we assume that Searle knows anything about whether the
> > Chinese Room understands chinese? Searle is merely an active component
> > in the chinese room's operation, but not necessarily the component that
> > must have awareness of its understanding. The person to ask is the
> > chinese room itself. Presumably, it would answer "yes, of course" .. in
> > chinese.
>
> SH: Ah me, the "systems reply" again, in its standard incarnation. Yes I
> gave the answer in the morning lecture [at MIT]; so did Searle, in his
> paper. But one has to be prepared to draw the logical consequences for
> one's beliefs, and the systems-repliers cling too strongly to their
> original premise to let go of it in the face of counterevidence.
> They just keep resurrecting it with more and more counterfactual
> science fiction whenever it is cast in doubt. >
> Here we go again: >
> The right version of Searle's Argument to focus on is the one where he
> has memorized all the symbols and all the rules for manipulating them.
> Then it's clear that there's no one to point to but Searle himself (so
> forget about the mind of the room). Then the only recourse for the
> systems-replier is that there must therefore be two minds in Searle,

> dissociated from one another: an English one, and a Chinese one (that's
> the second "system"). The simple answer to this is that although there is
> clinical evidence for the existence of multiple personality disorder --
> two or more minds co-existing in the same brain, with one not knowing
> about the other -- there is unfortunately no evidence whatsoever
> (nor is there ever likely to be) that one of the possible etiologies
> for this disorder is memorizing a bunch of symbols. Early sexual abuse
> maybe, but not symbols. To believe otherwise is pure sci fi.
>
> Stevan Harnad
>
> Okay, the logical point arises here. "There is no evidence whatsoever..."
> Quite true. But the Searle argument is a reductio ad absurdum, made in
> the context of an assumption (to be shown absurd). That assumption is,
> "Strong AI is true." In that context, we need no further evidence of
> a second personality inhabiting Searle's hardware. We've made an
> incredibly strong (and useful) assumption, that running the appropriate
> program would create the personality.
>
> I could go on, but let me pause to ask if you're following my argument?

I follow your argument, and will respond when you've expanded it into something big enough for me to sink my teeth into. On the face of it, the "incredibly strong (and useful) assumption" of Strong AI was that THINKING = SYMBOL CRUNCHING. The next step was Searle, showing this couldn't be so. What was the next step? A reiteration of the original assumption (T = SC), louder, explaining away Searle's counterevidence by further, ever more far-fetched assumptions (memorizing symbols creates multiple personality). That sounds like straightforward theory-saving, worthy of Ptolemy and the flat-earthers, and their respective strong assumptions...

> Well, I'm afraid that if I amplify you will focus on the amplification
> and not the original point. Isn't there enough to chomp on now? My
> observation seems clear enough to me: if we assume strong AI is true,
> then there exists a program the execution of which will bring a person
> into being. No further evidence required; it's merely a restatement
> of the strong-AI position

Look, there's a little thing called the other-minds problem: There's no way anyone can know for sure that anyone but himself has a mind. You're free to suppose either way about any other entity but yourself, including stones, thermostats, computers, plants, animals, people, heavans and gods. Not even Searle can "prove" that that computer there, whose program he is duplicating, doesn't understand Chinese. Maybe HE doesn't yet IT does. Who's to say? (There's Strong AI's postulate that EVERY implementation MUST have the property, but lets pass over that clause for now, because it's exactly the one that makes Strong AI vulnerable to Searle's unique little peek into the other-minds problem.) In fact, for the very same reason, maybe there are two minds in Searle, and one doesn't know about the other. Who knows?

Now that's old familiar stuff. What's new? It seems that Strong AI has decided to cash in on the other-minds problem by fiat, and without giving anything in exchange. Strong AI makes its point by Strong Assumption: We assume that running the right program gives you a mind, and, well, who's to say otherwise? You show me what looks like counterevidence, and I'll defeat it by strong assumption. Why? Because this assumption is USEFUL. Useful to whom, for what? Well, useful to those who assume it, of course, for they are free to believe as a consequence that all kinds of things think, understand, have minds. But we were free to do that all along! We could do it with stones and thunder. Why is it more useful to assume that symbol crunchers have minds than that they don't? And aren't you a little worried about making indefeasible assumptions? I mean, is a hypothesis to be abandoned only if an independent technique fails to have useful results? By that token, I can assume that computers are gods too, until they fail to be useful.

No, I think argument by Strong Assumption is a very bad strategy, and leads exactly to the hermeneutic circle that keeps spinning out of these exchanges. The critics challenge and give logical and empirical counterevidence to the assumption (such as the normal etiology of multiple personality) and the advocates just reaffirm the strength of their assumptions...

> Okay, I'll add a little more:
>
> (a) The program is not SAM. (The Churchlands made this point.) Even
> Schank never claimed that SAM would bring a person into a being.
> (Actually, I'm not sure Schank is a believer in strong AI.) If we
> knew how to write this program now, then the battle lines would be
> drawn quite differently. All we can do now is suppose that such a
> program can be written. But having supposed it, we can make use of
> that supposition.
>
> (b) By all means, be sure to connect the abstract symbols up with the
> world. For me, it would suffice to have the output squiggle-squoggles
> be buy and sell orders on the stock market, based on input
> squiggle-squoggles that were from today's Wall Street Journal (Beijing
> edition). Our virtual person could then have a Swiss bank account,
> and (provided you, Searle, and some government recognized his legal
> rights) get rich.
>
> But *please* don't focus on this part of the argument. I am not
> arguing against your theory of symbol grounding at this point, but
> against Searle's intuitions about consciousness and understanding.
> Searle himself always discounted the importance of connecting up the
> symbol system to sensors, so I just tossed this in to defuse the whole
> issue. The point I want you to respond to is the purely logical point
> about our ability to take advantage of the assumption we made at the
> outset.

Logically and methodologically, AI is free to write smart programs, including Swiss banking programs, that are useful, etc. For this freedom, there's no need whatsoever to assume that the programs have minds. Do do clever things they no more need to be mental than they need to be divine. AI can even try to write programs that pass the TT -- those would certainly be supremely useful -- still with no need whatsoever to assume they have minds. So what WORK does the assumption do?

I think it originally did three kinds of work, but now only two: Originally it was perhaps (1) a bona fide hypothesis (not an assumption) about the mind: But that hypothesis was refuted by Searle's thought experiment, and the refutation was explained by symbol grounding considerations. What was left? (2) A hermeneutical habit, that of freely adopting the "intentional stance" toward one's programs, interpreting them mentalistically; innocent enough, but irrelevant to the rest of us, and to science. Finally, let's admit it, where the promise of usefulness was not enough, the claim to have a line on how the mind works also probably served to (3) loosen some purse strings. Another criterion of "usefulness," perhaps...

> It seems to me that the Chinese Room argument has the same structure as
> the following argument that the earth is round:
>
> A. Suppose the earth were flat.
> B. Then, for crying out loud, the earth would be flat, and the earth's
> not flat!!!! Give me a break!
> C. (coda) I mean, what *evidence* would you have that the earth was
> flat???
>
> No self-respecting flat-earther is going to buy this, and no self-respecting
> computationalist is going to buy the Chinese Room argument either. (Which
> is the same, with "person = appropriate program" replacing "earth is flat.")
>
> -- Drew

Drew, surely you don't mean this. The hypothesis was T = SC, and as long as we considered only clever computers that couldn't tell us otherwise, there was evidence to support it. Then Searle pointed out that, according to this hypothesis, he ought to understand Chinese too if he memorized a bunch of meaningless symbols. The original hypothesis had not anticipated that; in fact, as long as the symbols were on the black-board, the "systems repliers" agreed that Searle wouldn't understand, but that Searle plus the blackboard would (the other-minds problem of course leaves room for that). But then when Searle pointed out that he could memorize the symbols, they replied that, well then, he would have TWO minds.

Hypotheses must take their consequences, both empirical and logical. T = SC has some pretty bad ones, not at all analogous to A - C above. I think you've picked the right kindred spirits for Strong AI (flat-earthers), and their reaction pattern is certainly as unregenerate, put the arguments confronting them have a bit more content than the ones you mentioned, more like: if the earth's flat, how come no one's ever fallen come to the edge? (Because it's infinite.) Then how come if you keep traveling in the same direction you end up where you started? (Because compasses deviate.) Etc. etc. Which just goes to show that you can overturn an empirical hypothesis, or at least force it into epicycles, but not a Strong Assumption.

Stevan Harnad

DON'T BLAME DESCARTES

> From: EFP@vms.cis.pitt.edu (Beth Preston)
>
> Here are a couple of comments on some things you said in your
> last posting. Sorry it's taken me so long to get around to
> producing them!
>
> ...we lack the innate empathy
> (knowing what it's like on the inside) that makes the TTT so sensitive
> with our own kind.
>
> I don't understand this at all. I thought the TTT was a behavioral
> test, and that that is why it counts as an objective test at all.
> So it seems like its sensitivity must depend on our knowing a lot
> in detail about how other people BEHAVE under various circumstances,
> not on any kind of empathy. If what you mean by empathy is saying
> to yourself things like "Yes, that's what I would have done if I
> were him" or "Yes, that's exactly what someone like her would do",
> then you are still construing empathy behaviorally--it doesn't have
> anything to do with vicariously feeling what it's like to BE that
> other person necessarily.

I disagree. The TTT has both a formal, objective component and an informal, subjective component. Cognitive and biobehavioral scientists will catalogue and try to generate all human performance capacities in robots. Once they're in the ballpark, there will be subtler (behavioral) cues that will be difficult to catalogue and characterize formally, but that people will be able to pick out intuitively. They might even be things like whether the candidate is moving or speaking in a humanlike or a machinelike way. Not all of these will MATTER to the human observers' ultimate judgment of whether or not the candidate has a mind, but they will play a potential role, and I doubt that there will be a way of judging them entirely objectively, i.e., obesrver-independently. And, yes, those cues will be based on our ability to vicariously feel what it's like to be in the kind of mental state someone else is in.

> And the TTT is as conclusive as it gets, not because of "total sensory
> involvement," but because it's as much as we can ask for and as much
> as we ever get by way of objective evidence of anyone's having a mind.
>
> I think I agree with this (under some construal), but it does not
> follow that the other minds problem is insoluble. You can take
> the TTT as being just what it MEANS to know that something else has
> a mind. See below.

What it means to know something else has a mind, and what it is for something else to have a mind are not the same thing. The TTT may be my only way of knowing, but since it's no guarantor, it's no solution. We must keep epistemic and ontic questions separate, as usual.

> And I certainly can say the other-minds problem is insoluble, for
> irrespective of what "definitions" or "criteria" I adopt, the only way
> to be sure anything has a mind is to be that thing, and I have that
> privilege only in one single case: my own. All others are separated by
> an impenetrable barrier. That's the insoluble other-minds problem.
>
> Yes, but you can just as well say that the other minds problem is a
> pseudo-problem. I.e., it never really was a problem at all until
> some philosophers (well, all right, Descartes--although Martin
> Luther seems to have had a hand in it too) invented an epistemology
> and an ontology which made it look like a problem. Give up that epistemology
> and/or that ontology and it just isn't a problem anymore, so it doesn't
> even make sense to ask about its solubility.
>
> Specifically, the insistence that the only way to know something
> else has a mind is to be that thing depends on a construal of
> "know" as certainty based on direct acquaintance.
> (It is well to remember that this epistemology also led
> directly to the external world problem, i.e., the assertion that
> you couldn't know whether there was one or not. I do hope this
> seems as patently absurd to you it as does to me.) But anyway, in
> the case of other minds, if you insist on direct acquaintance
> and you have the sort of dualistic ontology which makes the
> mental a private realm to which each thinker has privileged
> access, then and only then do you get the classical problem
> of other minds. I grant you that it is insoluble under
> the Cartesian assumptions. But it isn't insoluble per se,
> since under other assumptions it isn't even a problem.
>
>
> Beth Preston

Nothing to do with Cartesian assumptions. And plenty of people articulate it without knowing Descartes or philosophy or anything. It's the same as the question of whether an animal really feels pain if I pinch it. That's a perfectly well-framed, contentful question; it calls for and draws on no philosophy (at least no formal training therein); it admits of either a "yes" or "no" answer, both coherent, and indeed one true and one false in any given case, i.e., there's a fact of the matter. And no certainty (except if you're the one being pinched). We've been hearing, and swallowing, this nonsense about the insidious effects of the "Cartesian" legacy for too long; it's like another bad Whorf hypothesis about how a Cartesian education shapes our view of reality; it's a much less sophisticated issue than that, and open to the ordinary folk musings of any untutored (and unprompted) New Guinea bushman. And robotics is up against it in spades.

And of course there's an external world problem; everyone who has wondered about whether something's really happening or just a dream or hallucination knows that, though I imagine that fewer people have reflected on its implications. It's not a scientific issue, though, the way the other-minds problem is, except as another manifestation of the mind/body problem, which I also take to be quite real and not just a Cartesian facon-de-parler.

Stevan Harnad

> Date: Tue, 20 Feb 90 21:44:25 EST
> From: B Chandrasekaran b_chandrasekaran@cis.ohio-state.edu
> To: harnad@Princeton.EDU (Stevan Harnad)
> Subject: Re: Thinking as having images
>
> At this point, I haven’t given any thought to where it is supposed to go.
> I just meant to send it to you for your comments, since it is on an issue
> quite close to what you have been talking about, but builds on an earlier
> theory of mine concerning images. I am enclosing the draft here.
> ---

> Thinking As Imaging: How Semantics Gets Into The Act in a Symbol Processing
> Machine
>
> Outline of argument:
>
> For this discussion, without loss of generality (I hope), let us assume that
> the Chinese text is restricted to discourse about the spatial world,
> say about objects and shapes of things in the scene.
>
> According to Searle, and many workers in AI itself, the ‘‘strong-AI’’
> assumption is that thought, hence e.g., translation from Chinese to
> English, is execution of an algorithm. Let us say that such a Turing
> machine (TM) goes thru state transitions purely driven by ‘‘syntactic’’
> rules about symbol manipulation. Searle’s point is that when he goes
> thru those state changes and finally arrives at the final answer, he
> may be producing the correct answer, but he knows he is not
> understanding. Unless one holds that one may in fact be in a state of
> understanding without being aware that one is, Searle seems right about
> his mental state, in the ordinary sense of the term ‘‘understand.’’
> I also agree that the ‘‘systems reply’’ is not responsive to this issue
> either.

Two misunderstandings: (1) Strong AI does not refer to "translation" between anything and anything. It is the theory that thinking simply IS symbols and symbol manipulations. (2) There CAN be understanding without awareness of the understanding, but only in a system that has awareness and understanding in the first place; moreover, understanding without awareness does not make much sense when it is understanding without awareness of a whole language, except perhaps in glossalalia and trance; but there is no evidence that memorizing and manipulating a bunch of meaningless symbols and rules for manipulating them as Searle does (in his thought experiment) can induce glossalalia or trance.

> Knowing something about the structure of languages and how translation
> programs work, it is reasonable to hypothesize that there is a certain
> functional architecture to the set of instructions that Searle is using

> for translation, i.e., there is an internal structure to the program
> which mirrors the meaning structure. Specifically, at appropriate
> points in the evolution of the computation, intermediate symbol
> structures are created which correspond to the semantics of the
> situation: objects, relations, objects-in-relations, etc.

To repeat: Translation is not at issue; and symbols, be they ever so intermediate and ever so interpretable are still just meaningless symbols. No argument to the contrary has been provided by anyone; and to project an interpretation on them is just to beg the question (and lose oneself in the hermeneutic hall of mirrors).

> Now, imagine
> that the rule book is modified such that for each intermediate symbol
> or symbol structure on the right hand side of a rule that is executed,
> there is an additional set of instructions which describe the
> corresponding picture so Searle can imagine it. Now as Searle is
> simulating the algorithm, he not only is going thru the TM states, but
> he is also seeing images that come together into representations of
> objects, their shapes and their relations that the sentence that is
> being translated stands for. His performance in translation is
> identical, but Searle is now having some of the phenomenology of
> understanding that he didn't have in i above. He will have difficulty
> now in flatly asserting that he doesn't understand Chinese.

Yes indeed, but unfortunately all these extra stipulations simply change the subject and beg the question! The question was whether symbol crunching alone equals understanding. The answer is no. Giving Searle additional instructions (in English, presumably) that describe images, or the images of what the Chinese refers to, is simply irrelevant to the question at issue. Once you have Searle doing things the computer (a pure symbol cruncher) isn't doing, all bets are off. In fact, you may as well give him the English translation without fooling around with "images," since it's clear that the thought experiment no longer addresses any question of interest: Who cares what PEOPLE (who we know can understand) do when they crunch symbols AND are told what the symbols mean!

> But there
> is a problem: the images have no causal effect on the sequence of TM
> states that are being produced. That seems to run counter to the
> phenomenology of understanding. We think understanding something should
> have effects in our mental behavior in the future, i.e., understanding
> is not an epiphenomenon. If understanding in the sense of having the
> right sequence of images has to play a causal role, we need to modify
> the assumptions about the rule book even more.

There may very well be images in my head when I understand. But inducing them in SEARLE while he's in the Chinese room unfortunately shows nothing about anything, not even about whether or not images are epiphenomenal. It's just an arbitrary variant on what was previously a coherent and rather decisive test of a hypothesis (the symbolic theory of mind); the variant is just a Rorschach image, open to anyone's interpretation, neither testing nor deciding any particular hypothesis.

> Suppose the following system is adopted. Whenever an image is presented
> to Searle by the rule book, there is also another set of instructions
> which tell him to extract information by performing certain operations
> on the image (without loss of generality, I am assuming that all images
> in this discussion are visual). Suppose further that the rules in the
> rule book are such that the next state of the TM is a function of the
> current state of the TM and information from the image obtained as
> indicated above. Thus, the evolution of states of the TM that Searle is
> ‘‘simulating’’ is now causally affected by Searle having the image, but
> it is no longer just the TM that is producing the translation. At the
> very least, this is now a complex machine which has a general purpose
> TM embedded in an image-maker and image-reader.

Nothing of the sort, it seems to me: You’ve got Searle doing symbol crunching (simulating the pure symbol crunching computer, "Charlie," which is hypothetically passing the Chinese TT) AND you’re describing images to Searle (in English, presumably). Now you’re suggesting that Searle’s states (symbolic or otherwise) are influenced not only by the Charlie-simulation, which is ALL symbol crunching, but by Searle’s images -- which is undoubtedly true, but completely irrelevant, since CHARLIE isn’t conjuring up any images (and can no more understand English in this respect than Chinese)! It’s no longer clear who’s supposed to be simulating what, and what’s at issue. You’re just doing hermeneutics (projecting interpretations) on images in heads.

> I will shortly make a proposal about how the image representation and
> manipulation can also be viewed as information processing, but the
> medium in which the symbols reside and interpretation takes place
> retains some essential aspects of the semantics of the domain. This
> special purpose architecture gives the explicit connection with
> semantics that Searle is looking for. This is in contrast to
> general-purpose TM’s whose generality *requires* that their
> architectures are not biased to the semantics of any specific domain.

So far, however, you haven’t said what images are, where they come from, and what they have to do with symbol crunching, and why. You’ve just added them by capitalizing on the true but utterly uninformative fact that when you decsribe something to people who understand the language you are describing it in, those people are usually capable of seeing images of what you’re describing. So what? And how does that help to decide whether or not mental processes are symbolic, and if not symbolic, what they are instead?

> Before I proceed to a description of that proposal, I would like to
> dispose of two issues, one relating to whether connectionist proposals
> for mind avoid the problems with the original version of the Chinese
> Room, and the other about whether the notion of images that I just
> talked about is either necessary or sufficient for understanding.

Unless I have missed something, you have given no notion of images other than the phenomenological one we had all along, and that’s no help.

> Taking the question of connectionism first, in my view, connectionist
> architectures are just as subject to the problems pure
> “symbol-crunchers” face with respect to semantics. As I just
> discussed, what makes a general purpose TM universal is that the
> machine’s architecture is not biased towards the semantics of any
> domain, modality or world: the semantic properties of the domain that
> is being simulated are realized by the functional properities of the
> “software”, the set of instructions. Thus watching a Turing Machine
> in operation at the level of the hardware, all one sees are symbol
> crunching, whether the symbols “stand for” Anna Karenina or an air
> molecule.

True; but what other "level" does a TM operate at? It is, after all, a pure symbol cruncher. The symbol crunching is semantically interpretable, to be sure, but how to GROUND that semantic interpretation in some way that is not merely parasitic on the ideas in our heads is the problem at hand. Until further arguments are provided, the answer is that pure symbol crunchers are simply ungrounded, and that it’s not just a "level" problem. Taking something else, like a person, and point out that he, like every other system, is probably "equivalent" to a TM does not mean that he IS a TM, any more than an airplane or a furnace is a TM. In fact, this just goes to show that the matter of "levels" always gets inverted in the "hermeneutic hall of mirrors" (in which this kind of argument is hopelessly lost, in my view). For it is nonsymbolic systems, such as real, understanding persons, that always have, among other things, a semantically interpretable but purely symbolic, TM "level" of description, whereas you are seeing it the other way round: As if symbol systems always have a nonsymbolic, unproblematically semantic "level." There is no evidence whatsoever that this is so. They are merely INTERPRETABLE that way. That interpretation must be grounded, not taken for granted.

> Connectionist machines, qua general architectures to model vision,
> speech processing, language understanding, etc., also necessarily
> achieve this generality by offering a general “language” in which the
> phenomena of information processing in these various domains can be
> encoded. A connectionist machine, in operation, is “just weighting,
> summing and thresholding” exactly as much as TM operations are “just
> symbol crunching”. This is not to argue for TM style information
> processing versus the connectionist one as the “right” style, but to
> say that for the issues involved in the CR, specifically the problem of
> how semantics of the world are brought in during information processing
> of a syntactic nature, connectionism per se does not offer a royal road
> for resolution of the puzzles. Additionally, my image proposal, even
> though it will be couched in the symbol processing language, is not
> dependent on it for its essential point.

I have not yet seen an image proposal, just the projection of the "image" image (sic) in the hermeneutic hall of mirrors. It’s just as easy to project images as to project meanings. The nontrivial part is to ground those projections in something other than the interpretation of otherwise meaningless, imageless symbol crunching (which is simply parasitic on OUR images and meanings).

I agree that there seems to be little difference between the symbol crunching of a TM and the weight-summing of a net. But one critical difference is that symbol crunching is, for better or worse, wedded to the implementation-independence postulate of Strong AI that I have dubbed "teleportability": If symbol system S is the right symbol system for capturing property P (say, understanding Chinese), then EVERY implementation of S must have property P. P is "teleportable" into every implementation. This is why Searle, who can implement S, can give us the bad news: that S hasn't captured P after all. Teleportability allows this one narrow periscope through the otherwise impenetrable "other minds" problem.

Now weight-summing nets may be no more capable of understanding than symbol crunchers (or stones, for that matter), but unless Searle can implement them, he can't say it is or isn't so. (As I've written before, however, he CAN apply the Chinese Room Argument to a symbolically simulated net, as opposed to one that is essentially parallel; and unless someone can prove that there is something essential about the parallelness for capturing understanding, that's almost as good as the case against pure symbol crunching.)

> The second issue has to do with the relationship of the phenomenology
> of human understanding with the essence of what is being understood.
> The problem with the ‘‘understanding as having images’’ view that I just
> offered is that it makes understanding completely dependent upon the
> specifics of the human perceptual systems. While of course the specific of the
> human perceptual systems will limit to some degree what can be understood,
> there is also the sense of understanding as transcending these limits.

The real problem, it seems to me, is that you haven't shown what images are, and how they are related to symbols.

> In order to answer this question, we need to seek the relationship
> between the image and the object of understanding. What Searle is
> building up in his head when he is understanding is, among other
> things, a 3-d model of the shapes of objects in the scene that is being
> described and the visual image only partially reflects this underlying
> reality. Kinesthetic ‘‘images’’ contribute to and are invoked by the
> underlying 3-d model of the world as well. The underlying ‘‘image’’,
> the 3-d shape model of the world, is in many cases available by an
> integration of the visual images in various views as well as from the
> kinesthetic image. Thus operations on the images in various views give
> the agent de facto capability of operations on the 3-d model of the
> scene, and hence of acquiring information about the world.

There may or may not be something to this; it is hard to say, because you have not yet said what these sensory images are. What is certain, however, is that it is adding nothing to bring Searle and the Chinese Room into any of this. You might as well talk about any of us, and the unexplained images and meanings we know THAT we all have, but not HOW! Searle's attempt to implement a symbol cruncher to invalidate Strong AI simply seems to be a red herring in all of this (and vice versa).

> The effectiveness of some forms of thought -- deductive reasoning, e.g.
> -- depends on the fact that the symbols are separated from their
> meanings, so that the form of reasoning can be checked without
> interference from the meanings of the symbols. Here the understanding
> that is involved is of the form of reasoning itself, not the referents
> of the symbols. But this is really a rather rare form of thought and is
> not representative of thought in general. In general, thought proceeds
> concretely by evocation of images and their interaction with symbol
> manipulation.

I can’t understand your imagery here. And though I have an idea what symbols and symbol manipulation are, I still don’t know what their meanings are. Hence I am no closer to knowing what thinking and understanding and images are.

> It is often asserted that a simulation of thought is still thought, and
> in this it is unlike a simulation of typhoons. We need to be more
> precise here: there *are* some forms of simulation of thought that
> remain thoughts, while others, which do not invoke and causally use
> appropriate images *serve* some of the functions of thought, without
> necessarily being thoughts.

I can’t follow this at all (and it seems to me that you mean to say the opposite). I assume that a simulation is a symbolic simulation. Again, I know what symbols and symbol manipulation are. So the simulated vs. real thought/typhoon question is really whether thoughts/typhoons are just symbols, crunching. The answer is no, in both cases. So what are images? They can’t just be symbols crunching either, for the same reasons; but until you say what they are, and how they relate to symbols and symbol crunching, you are merely giving us images rather than an explanation!

> At this point, it may be said that my acceptance of the central
> importance of image invocation as well their playing a causal role in
> thinking is pretty much giving away the store to Searle: having and
> using images seems like the antithesis of the mainstream of AI work,
> which is pretty content with image-less symbol manipulation, and
> doesn’t see a need for it. The traditional AI assumption about how
> sensors and bodies get into the picture is that sensors give
> information which is encoded as symbol structures, which are
> manipulated by thought processes, and certain conclusions are reached,
> which, if they relate to action, are transmitted to action transducers
> to get the action accomplished. In this view, there is not much of a
> role for images as causal participants in thinking. However, the tide
> is turning: the work of Chapman in planning is beginning to look at how
> visual representations of the right sort make an enormous difference in
> the complexity of planning as a problem solving activity. In our own
> Lab, we have been looking at image-driven naive physics reasoning --
> e.g., how forces on some objects in the scene will play out. Here the
> images again play a causal role since the inference rules are no longer
> symbols to symbols, but abstract images to abstract images. Reasoning
> here consists of image-specific operations in which the specific scene

> is matched against abstract images, a new abstract image is inferred
> using the inference rule, this rule is instantiated to the
> scene-specific image which is the next step in the qualitative
> simulation of the situation.

Sounds interesting, but what are "abstract images"? If they're something other than symbols, then there's hope for grounding here (but then what are they?); if they're just symbols INTERPRETED as images, then this is just symbolic AI, dressed up in yet another hermeneutic guise.

If really you reject (as I do) symbolic AI's haste to get from transduction into symbols, and its belief that it is in the symbol manipulation that the real work is being done, you will have to be careful to make sure that your "images" are not just symbols that are interpretable as images; and that the constraints that supposedly come from the images are really nonsymbolic constraints, and not just the usual symbolic constraints INTERPRETED as nonsymbolic.

> In what remains I want to give an outline of a theory of image
> representation that has roots in Marr's work on vision, and it arose in
> my attempts to reconcile the symbolic versus analog debate about
> imagery. (There is no need to rehash that debate here.)
>
> Let us assume, for the sake of simplicity, that the function of the
> early visual system is to interpret the scene in terms of 3-dimensional
> shapes. (Recognition and labeling come after.) We need a vocabulary of
> primitive shapes in terms of which an arbitrary shape can be
> described. There have been several alternatives proposed, and for my
> purposes it does not matter which one is used. I will follow Marr in
> the use of a family of generalized cylinders for such a
> representation: any shape can be described as a hierarchical
> organization of appropriate cylinders chosen from this family. The
> cylinders are parametrized by length and diameter. The description
> language includes terms which specify how the primitive cylinders are
> attached to each other: subparts and their orientations. Each level of
> the hierarchy corresponds to a description of the scene at some level
> of resolution. For example, Marr has a description of a human figure
> where at the top level, the scene is just one cylinder.
> At the next level of the hierarchy, parametrized cylinders that describe
> the face, the torso, the arms and the legs are introduced. This is continued
> until the image is represented to the degree of detail desired.

So images are just symbolic descriptions that are interpretable as descriptions of geometric shapes?

> Let us call the description of a scene in this language of primitive
> shapes and their relations the image symbol structure (ISS). ISS can
> also be thought of as a set of instructions, an algorithm, to construct
> the scene, but the interpreter which can use the algorithm is a special
> purpose machine. You can think of a robot with access to bins where are
> stored cylinders of various sizes. The robot interprets the shape
> symbols (with parameters attached to them) in the ISS as instructions

> to fetch cylinders of appropriate size from one of the bins, and the
> relational symbols (again with parameters that describe the relative
> locations of the constituents and their orientations) as specifying how
> to put the cylinders together in 3 dimensions.

Are we talking about real robots in real worlds here, or symbolically simulated robots in simulated worlds? I have nothing against symbolic simulation, of course, if it can be cashed into the real thing. But if the world is so impoverished as to be only a stylized toy world, and the only way the symbols are cashed in is through our interpretations, then we may still be trapped in the hermeneutic circle.

> An image of the scene on the retina results in the ISS being produced
> as in the above analysis. Conversely, I propose that the ISS is the key
> to how a mental image can be represented symbolically but can be
> perceived internally as if it were a picture. Imagining or remembering
> an image involves the corresponding ISS stimulating the perceptual
> machinery, much as the robot above constructed the scene from the
> instructions contained in the ISS. The symbols in the ISS are not being
> interpreted by a general purpose TM, but a special purpose machine
> whose primitive operations are specialized to the image domain, i.e.,
> they preserve the needed semantics. Watching this special purpose
> symbol processor at work, one does not merely see ‘‘symbol-crunching’’
> at work, one sees the symbols being interpreted in a matrix of
> meanings.

"Matrix of meanings" is just a figure of speech. All I see here is the headlong rush into symbols that I spoke about above. It’s not at all clear what causal work the nonsymbolic structures and processes (e.g., transduction) are actually doing, and how, and why. As in standard symbolic AI, the picture seems to be one of symbol crunching doing all the real work, and somehow managing to be "connected" to sensory input and the outside world "in the right way." That’s not grounding; it’s blind faith, bolstered by a lot of interpretation. (How are semantics "preserved"? And how are the symbols "interpreted"? What grounds ther interpretation?)

> I need to sketch how the ISS, which at this stage is merely a shape
> representation, is used for recognition and how it is integrated with
> other forms of symbols, specifically conceptual symbols. Recognition is
> the association of labels with abstract forms of ISS -- abstract in the
> sense that ISS’s of specific instances are abstracted into an ISS that
> captures the essence of the recognition concept. The ISS corresponding
> to such a recognition concept is still a representation of an image,
> and now will invoke a generic image for the concept in question. We
> also need mechanisms by which a phrase such as ‘‘an elephant eating an
> apple’’ is converted into a composition of constituent ISS’s into one
> that stands for the phrase as a whole, and for how that ISS in turn
> evokes the image.

All the real work here seems to be that of symbols, crunching; the rest seems to be just interpretation. Why/how is a shape (or image) representation really a representation OF that shape (image), rather than merely being systematically interpretable that way by us?

> Now let us see how this view of the image representation -- from a real
> image to the creation of an ISS and from the ISS to the invocation of a
> mental image -- helps in the CR problem. What Searle was able to do
> with the images in front of him in the modified CR scenario was to
> perform some operations using the semantic properties of the image.
> To the extent that the invocation of the mental image is done using the
> ISS, which in turn is being constructed from the symbol structure
> describing the scene in the instruction book, we still have an
> information processing account, but now it is integrated with a form of
> grounding in images.

I don't see any grounding in images here. I just see symbol structures being interpreted as being images and as being connected to sensory input. If at least I saw how the shape of the sensory input influences the symbol crunching I might get an idea of whether any true grounding could be going on here. But from what I understand of what I have read here, however, the sensory input is just dangling there, mysteriously linked to the symbols in the right way. (What is a "real" image, by the way? and a mental one?)

> I have assumed a rather limited domain of discourse in the above
> argument: one of describing phenomena relating to objects and their
> shapes in the spatial world. One might ask: how about purely abstract
> concepts, like justice or nutrition? I am told that Wittgenstein has
> argued against the image theory of meaning that my account strongly
> assumes. I haven't worked this all out, but it seems to me that all
> understanding is ‘‘grounded’’, to use Harnad's term, in perceptual and
> kinesthetic modalities.

So it might be, but the question is still, "how?" In my Symbol Grounding paper I stuck my neck out and said specifically that the analog shape of the sensory projections and of the sensory invariants that reliably pick out categories of objects constrain (ground) the NAMES of those objects; the names then serve as the grounded elementary symbols in a hybrid, bottom-up nonsymbolic/symbolic system. It is hybrid because, unlike in a pure symbol system, the symbol crunching is NOT governed exclusively by the arbitrary shape of the symbol tokens themselves, but also by the NONarbitrary "shape" of the sensory icons and invariants in which they are grounded. That's how my grounding scheme would work, but from what you have described, I cannot discern in what the "grounding" in your proposal would actually consist.

> Understanding, in the sense that Searle wants to give it, is a property
> of a deliberating agent. The account I have given above, which I think
> has close parallels to the accounts given by Harnad, indicates how the
> symbols in deliberation are connected to corresponding symbols in
> different perceptual modalities such that all acts of deliberation are
> ipso facto acts of imagining as well. Since the imaginings take place
> in media which preserve the semantics of the underlying world in some
> relevant dimensions, and since the operations of deliberation are
> partly operations on these images, ultimately, the act of thinking is
> guided by images of the world, which is how semantics come into the
> picture. When one does pure symbol crunching, one may often achieve the
> identical effect locally by capturing the functionality of the

> semantics in the functional architecture of the software, but it is
> ultimately limited since the general purpose Turing Machines get
> increasingly bad at such simulations, bad not in the sense of
> computability, but complexity, learnability, etc. On the other hand, my
> account shows that the problem of understanding as posed by the CR
> problem is not fundamentally beyond the capability of machines.

I'm afraid I don't know what you mean by "deliberating" (except in the robotic sense of acting in the world in a way that is constrained by the consequences of the actions in some regulatory way). I know even less what you mean by "imagining." What *I* would mean by imagining is the activation of sensory icons and invariants, and of their connections with primitive and composite symbols. I also don't know what you mean by "preserved semantics." I know that pure symbol crunchers have only the semantics projected on them by people with real semantics in their minds. My own candidate for that real semantics is grounded hybrid systems of the kind I've proposed, which can pick out the things their symbols refer to, whereas I am regrettably unable discern what your candidate is (on the basis of what I have read here)

Operations on "images" in my sense would be operations on sensory icons (which preserve the shapes of the sensory projections of objects and states of affairs), on their sensory invariants (which allow the system to pick out and name those objects and states of affairs) and on the symbols to which the icons and invariants are connected. For me, "thinking is guided by images of the world" because symbols are grounded in nonsymbolic icons and invariants. Pure symbol crunching does not and cannot achieve the "identical" effect, except in the TM sense of being INTERPRETABLE as images and thinking (just as it can be interpreted as flying or heating, without really flying or heating).

I don't think either complexity or learnability is what distinguishes simulation from reality in the case of thinking; I think Searle's CR Argument shows that pure symbol crunching simply isn't thinking, and my Symbol Grounding Problem shows the reason why. Finally, the problem is not whether thinking is beyond the capability of machines in general, but whether it is beyond the capability of pure symbol crunching machines.

Stevan Harnad

Date: Tue, 20 Mar 1990 17:17:44 EST From: B Chandrasekaran b_chandrasekaran@cis.ohio-state.edu

Stevan,

I have read thru your response once with some care, but clearly I need to go over it again. At this point I see that there have been a number of misunderstandings, largely caused by the terseness of the outline, by starting with the CR problem rather than the image proposal, by not giving hard details about the image proposal, and partly caused by your misunderstanding of my intent.

The last item first: In most of the outline, my intent was neither to agree nor disagree with Searle, but to try understand the role of images in understanding, since I believe that perceptual and motor images are the major carriers of 'semantics' in our heads. You have taken my entire outline as yet another modified-strong-Ai defense of symbol crunching. I say, "Strong AI assumption is that thought, hence e.g., translation, ^^^^^^^^^^^^^^^^ is execution of an algorithm." You say, "Strong

AI does not refer to "translation" between anything and anything .." Precisely. All I said was that translation is only an example of a task accomplished by an intelligence. If I had said, "e.g., making jokes", you might have said that "Strong AI does not refer to "making jokes."" I used translation as an example, since that fits in naturally with the Chinese Room specs. You go on to say, "..beg the question! The question was whether symbol crunching alone equals understanding." No that was NOT the question for me in the outline. I am quite willing to agree, and I thought that was the point of my outline, that symbol crunching alone does NOT equal understanding.

Let me simply repeat my claim about what i think is going on when people understand say a sentence. I think they have images, the images preserve the semantics of the world because the images themselves are not pure symbol structures, but are caused by, as you put it (as I equivalently put it in the larger paper on my image theory) , "the names serve as the grounded elementary symbols in a hybrid, botom-up nonsymbolic/symbolic system." The symbols in my ISS are the names (they are not labels yet, but names of shape primitives, but that shouldn't matter), the symbol structure invokes "the NONarbitrary shape of the sensory icons.." Hey, that is fine with me, and I thought that was what I was trying to tell you in the outline. Clearly it didn't come through. The full paper clearly lays out what the primtive icons are, how the symbol structure invokes the icons and builds a real image in an visual array.

Your complaints about the last para re deliberation etc are simply the result of overcompaction -- that was precisely why I was not posting it.

Finally, to repeat, I have no problem with the assertion that "pure" symbol crunching isn't thinking. I was also trying to show what kind of image representation framework can be constructed for machines such that it is not "pure" symbol crunching. It is clear that the outline didn't do it.

Chandra

From harnad Sun May 13 23:11:38 1990 To: dietrich@bingvaxu.cc.binghamton.edu Subject: Somnoloquy, Necrolalia, etc. Cc: harnad

Reviving the Symbol grounding discussion:

It's time to start again; I have a small backlog I will be replying to soon. Let me know if you wish to be removed from the list. -- SH

> From: Eric Dietrich
>
> Dear Stevan:
>
> In preparing for the SPP conference, I re-read your commentary
> "Computational Hermeneutics." Some thoughts I initially had months
> ago are clearer now. Here they are.
>
> 1. It seems to me that your criticism of computationalism
> rests on a particular metaphysics and a particular philosophy of

> science. If that is so, then it is no wonder that we seem to talk
> past one another: we have different fundamental assumptions.
>
> You say "We must distinguish between what can be described or
> interpreted as X and what really is X (p. 1)." Two paragraphs down
> you mention natural kinds (p. 1). These are two things I simply don’t
> countenance.

I don’t make much of natural kinds, so I don’t think we’d have any problem there. But what is it "not to countenance" the real vs. as-if distinction? I’m asleep and snoring, and my ZZZZ can be interpreted as meaning "My Country ’Tis of Thee," but I either really mean that by ZZZZ or I don’t. Isn’t it reasonable to say there’s a difference there?

> You talk as if there were only one reasonable position
> to take regarding "reality." And part of that position entails that
> there are natural kinds. I think "reality" is constructed by the
> interaction of human sense organs, stuff out there, and internal
> mental states (c.f. Kant). I, therefore, deny that there are natural
> kinds, and that there is a real distinction between interpreting
> something as X and that thing really being X. (I suspect Searle,
> too, buys into the same realist metaphysics and notions of science.)

I repeat, I don’t make much of natural kinds, and I think the kind of approximationism I advocate in the last chapter of "Categorical Perception: The Groundwork of Cognition" is as constructivist as you would like. In fact, I go further and say psychologists have no business doing ontology (pronouncing themselves on "what there is"); they should stick to explaining what the organism can do in the world. But NONE of this touches the distinction between mere interpretability-as-if-X and X: Just reread the stuff about ZZZZ above (or SSSS below).

> Your view therefore presupposes, it seems to me, an entire
> view of the world and knowledge which can be rationally denied.
> Therefore your approach to the Chinese Room Problem is merely one
> approach. It has the virtue of being consistent with your
> philosophical biases and assumptions. But then, so does my approach
> (which is essentially the systems reply).

Forget about philosophy. Suppose our goals are purely empirical. They may be (i) to build smart systems, irrespective of whether they’re smart in the human way, or they may be (ii) to explain the causal mechanism of human smartness. Either way, you’d better face the "X vs. interpretable-as-X-but-not-X" question. Otherwise, I argue, you will fail at (ii) or even at (i). The symbol grounding problem is independent of your choice of metaphysical poison.

> 2. Paraphrasing me, you say (p. 4) "Evaluating the expression
> ["SSSS"] requires the LVM to determine the value of the variable
> ["SSSS’"], which is some other Lisp expression, say, ["SSSS’’"]...
> etc. etc."
>
> I see nothing wrong with this reinterpretation of my quote. I
> don’t understand how it is supposed to cause me problems. It only

> seems denuded of its meaning because you have not allowed me to
> interpret the ["SSSS"] which I could do if I had a machine or some
> other system of interest in front of me. In fact, your ploy is a bit
> of slight of hand and question begging because the expression I used
> "(+ x 1)" has an interpretation which is associated with it which you
> simply ignore or refuse to believe exists: namely, x + 1. (The
> orthographic similarity is intentional: it facilitates reading,
> programming and debugging.) If the machine used ["SSSS"] it would
> still mean x + 1. You are pretending that that meaning isn't there.
> This seems fair to you because you assume natural kinds and
> realism. Seen you soon, Eric Dietrich

Of course "SSSS" is interpretable as "(+ x 1)," which in turn is interpretable as (+ x 1). The question is, when does it really MEAN (+ x 1), the way *I* mean it when I say it? (And I hereby disavow anything I may say in my sleep, or after my death...)

Best wishes, Stevan

From: Stevan Harnad To: oded%wisdom.weizmann.ac.il@pucc Subject: Re: As-if intentionality

Oded, you wrote:

> However, the distinction between intrinsic and as-if intentionality
> still bothers me. Why can't we treat all intentionalities as "as-if"?

We can "treat" all of them that way, but some will still be real and some not...

> The distinction, as one might say, is that "real" intentionality plays
> a causal role in the production of behavior.

No, at least that certainly wouldn't be my distinction. The difference between really believing/desiring/meaning that X and just being interpretable as-if believing (etc.) that X is that there's a subjective state corresponding to believing that X. If there's nobody home, there's no subjective state, and hence only as-if intentionality, regardless of causal role in behavior.

> But why can't we adopt a
> more mechanistic causal explanation of behavior, and treat explanations
> in intentional terms as metaphors? Why can't an intentional predicate
> be treated as approximating large (and paractically unexpressible)
> classes of physical states, as do chemical or biological predicates?
> There are some *mechanisms* in my body that make me open the door and
> make another cup of coffee. As side-effects, these mechanisms cause
> some states that correspond to what seems to be my desire/plan/etc.
> w.r.t. the coffee. But to say that these "representations" play a
> causal role in my behavior is an unjustified jump from correlation to
> causality. What can you say against this solipsistic(sp.?) approach
> towards one's own mental life?

I wouldn’t say anything against it, except to point out (a) that it’s epiphenomenalistic, not solipsistic, to doubt the causal role of volition (and I for one am happy to doubt it), but (b) this does not make intentionality one bit less real (or more "metaphorical"): To put it another way, the fact that my beliefs, desires, etc. do not have an independent causal power over and above whatever physical states/processes are implementing/causing those beliefs/desires does not mean that in some cases the beliefs and desires are not really there, and in other cases they are, or that the distinction is untenable, or merely behavioral. Realists care about what’s really the case; instrumentalists restrict themselves to what is empirically predicatable/explainable. It’s a free country...

> Let me ask you a personal question. What do you think about the
> academic world? As an inter-disciplinary type you have probably come
> across various communities, cults and churches. As an editor you have
> probably published (among other interesting and useful material) a lot
> of needlessly-lengthy articles/responses which had a very moderate
> contribution to anything at all, authored by respectable persons in
> their respectively respectable fields. What is your attitude toward
> this situation? Please excuse me for being so bold, and feel free to
> ignore this paragraph completely in your response if you feel this is
> an improper object for discussion.

I think most of the "fruits" of human endeavor -- be they practical, scholarly, scientific or artistic -- are of low quality, relative to the rare proportion that is of true value. That’s just Gaussian density/destiny, a biological fact. However, since it is not always clear which is which, and since a more rigorously selective civilization may not be feasible, we have to allow a large quantity of chaff in order to be ensured of finding the usual proportion of wheat.

Best wishes, Stevan

On the Reality of the Real/As-If Distinction

> From: Eric Dietrich
>
> I think I see the problem. Interpretation is constrained by various
> explanatory pressures. As I use the term, "interpretation" is
> scientific explanation applied to computational systems. So, it is
> false that your ZZZZ’s can be interpreted as "My country ’tis of thee."
> Interpretations have to be warranted by the behavior of the system and
> explanatory goals of the one doing the interpretation, and no behavior
> of yours warrants the interpretation of ZZZZ’s as a patriotic song.
> This is what the first sections of my paper "Computationalism" are
> about. There, I tried to show that computational interpretation is
> scientifically rigorous. esd

The point you are making is valid, but does not affect the distinction between real and merely interpretable-as-if: It is true that a symbol system, to BE a symbol system, must be systematically interpretable, not just arbitrarily interpretable. As I state explicitly in "The Symmbol Grounding Problem," anything will support an arbitrary interpretation, but only semantically interpretable formal systems will support a systematic interpretation. Fine. That was part of the DEFINITION of a symbol system in the first place. But now we have the problem of distinguishing bona fide symbol systems that can be given a systematic interpretation, such as the symbols in a book or those on your screen right now, from the symbols (if any) in my head and yours, which can not only be given a systematic interpretation as meaning something, but actually do mean that something. I mean MEAN in the way you mean it when you say "This is what I mean."

For example, when I utter the symbol "chair," it is not just that my utterances are systematically interpretable as meaning chair; they really mean chair, in a way that a screen (or a book) that tokens "chair" does not really mean chair -- because it doesn't mean anything at all; it is merely (systematically) interpretable as meaning something by someone who really can mean something. To put it Searle's way, the meanings of the symbols in my head are intrinsic to them, whereas the meanings of the symbols in a book or a computer are derivative -- parasitic on the meanings of the symbols in our heads. Or, as I would put it, they are ungrounded, whereas my symbols are grounded.

In what does the grounding consist? No one really knows what's going on in our heads, but Searle has at least shown that it can't be just symbol manipulation. And I've suggested some of the kinds of nonsymbolic processes it might be: The ones that allow us to discriminate, categorize and name objects on the basis of their sensory projections. That can't all be symbolic; transduction, for example, is not just symbolic. There's an essential difference between something that really transduces physical energy and something that is merely interpretable as if it transduced physical energy, just as there is an essential difference between something that really flies or heats and something that can merely be interpreted as flying and heating.

Enter the TTT. Unlike the TT, it's not just symbol manipulation and interpretation. Our robotic capacities are things we really do in and to the world, not just things we are interpretable as doing in the world. Now the fact that when I utter "chair" I can not only use the symbol in a systematically interpretable way, but I can also pick out the things it stands for, that fact doesn't guarantee that it has intrinsic meaning either (NOTHING can guarantee that), but it certainly gives much stronger grounds for confidence than just semantically interpretable symbol manipulations. And in fact, that's all you need to ground s symbol system -- this is where my "groundedness" parts company with Searle's "intrinsic intentionality": For even a TTT-passing robot may merely be acting exactly as if it really means what it says, whereas in reality it does not mean anything, because there's nobody home. (This is called the other-minds problem.)

So there are actually TWO real vs. as-if distinctions: one for the intrinsic vs. derived interpretations of symbols and the other for the mindful vs. mindless interpretations of robotic actions. Both are perfectly valid distinctions, and nothing is gained (and a lot lost) by pretending they're not there or not tenable. About the first we can say this much: In the case of heating, flying, and thinking at least, pure symbol systems are only as-if. (So much for "computationalism" as a theory of mind.) About the second all we can say is that the only way to distinguish a mindful TTT-scale candidate from a mindless one is by BEING the candidate. But that doesn't make even that distinction one bit less real.

Stevan Harnad

From: AMR@ibm.com (Alexis Manaster-Ramer [AMR]) To: srh@flash.bellcore.com (Stevan Harnad [SH]) Subject: On degrees of equivalence

This is a response to Stevan Harnad's comments on my remarks on Connectionists. My original remarks interspersed with the comments follow. Each comment is followed by my response. NB The original stuff has '>' in front of every line, Harnad's comments have no prefix, and my responses begin with **.

[SH: Subsequent interpolations from me are in square brackets, like this. SH.]

> AMR: (1) I am glad that SOME issues are getting clarified, to wit, I hope
> that everybody who has been confusing Harnad's arguments about
> grounding with other people's (perhaps Searle or Lakoff's) arguments
> about content, intentionality, and/or semantics will finally accept
> that there is a difference. >
> (2) I omitted to refer to Harnad's distinction between ordinary robots,
> ones that fail the Total Turing Test, and the theoretical ones that
> do pass the TTT, for two unrelated reasons. One was that I was not trying
> to present a complete account, merely to raise certain issues, clarify
> certain points, and answer certain objections that had arisen. The other
> was that I do not agree with Harnad on this issue, and that for a number
> of reasons. First, I believe that a Searlean argument is still possible
> even for a robot that passes the TTT.

SH: I'd like to hear that Argument! As I think I showed in "Minds, Machines and Searle," the robotic version of the Test is immune to Searle's Argument, which works only against pure symbol manipulation. The reason it works there is both symbol-manipulation's chief weakness and its chief strength. "Symbolic functionalists" ("Strong AI") call it "multiple realizability" and I lampoon it as "teleportability": It's basically the software/hardware distinction. Here's the argument, step by step:

The same program can be realized on many different hardwares. What matters is that in every implementation of the program the internal states stand in a certain functional relation to one another (and to inputs and outputs). Hence, any functional property a given formal program is claimed to have, ALL of its possible implementations must have that property too. So, in particular, if it is claimed that, say, understanding Chinese, is just symbol manipulation of the right kind, then once you commit yourself to the program in question (one that passes the Turing Test [TT] -- symbols-only version), EVERY IMPLEMENTATION OF IT MUST ALSO BE UNDERSTANDING CHINESE ("teleportability"). So, whereas the "other minds problem" normally presents an impenetrable barrier to ascertaining whether any other system than oneself is really thinking, in this special case, Searle himself can BE the implementation -- and can accordingly come back and report that, no, he did not understand Chinese in so doing, hence neither could that pure symbol-crunching computer (to which he is functionally equivalent, remember) have been understanding Chinese.

That's why the Argument works for the TT and pure symbolic functionalism. But, as I showed with my Transducer Counterargument (where the mental function under scrutiny happens to be, instead of understanding, seeing), it doesn't work for seeing, because seeing depends ESSENTIALLY on (at least one) nonsymbolic function, namely, transduction; and transduction, because it's nonsymbolic, either cannot be implemented by Searle (in which case the Chinese Room Argument cannot even be formulated, for Searle cannot duplicate all of the robot's functions, on pain of a VALID "Systems Reply," for a change) or he must BE the robot's transducers -- in which he case he would indeed be seeing. Now sensing and acting on the objects in the world are robotic functions. They are required by the TTT, but not the TT (at least not directly). Hence the TTT is immune to the Searlean argument.

** AMR: Three points. First, what I mean by a Searlean argument is ** weaker than what Searle said. I think he should have only ** argued that a standard computer program MIGHT possibly NOT ** have cognition even if it passes the Turing Test. He should ** not, I think, argue that it CANNOT possibly have it. The ** reason I say this is that Searle does not make it at all ** clear in what crucial respects a human mind differs from ** a program, so that he is in danger of contradicting himself. ** But I think that it would be reasonable to conclude that ** we may no longer agree with Turing that a program passing ** the (original) Turing Test necessarily ipso fact is intelligent ** in the relevant sense.

[SH: Oh my, where to begin on this one? Surely the Searlean argument is the one Searle made, and not the one you say he should have made, which he would in fact reject. Searle argued that a computer can't possibly understand JUST BECAUSE OF IMPLEMENTING A PROGRAM, because Searle himself could implement the same program and wouldn't understand. Of course the computer "might" understand anyway, but then Strong AI can't explain or take credit for that, because then that understanding couldn't be JUST because of implementing the program (what could be a clearer difference than that?). As to "necessity," the other-minds problem is always there to remind us that there is no way to ascertain that anything NECESSARILY has (or does not have) a mind except by BEING that thing. The TT certainly does no such thing, and never did, before or after Searle. Even Strong AI advocates take their position to be a hypothesis rather than a proof! So so much for "must" as opposed to "might." It is not even logically impossible that memorizing the program could cause Searle to have another mind that understands Chinese and that he is not aware he has -- but I think only those who have bought irretrievably into the epicycles dictated by the "systems reply" would want to hang on a possibility like that. Unless I have misunderstood, you do not appear to have grasped the logic of either side of the question. SH.]

** AMR: Second, if we assume that there exists a robot that passes ** the TTT and it is not a nontrivial robot (i.e., an artificially ** assembled human being), then I think that the brain-in-the- ** vat problem reasserts itself in a new way. Even if we grant ** Harnad's objections to the brain in the vat when applied to ** biological systems, I do not see how we could rule out the ** possibility of this robot being so structured that its ** "mind" can be separated easily from the hardware and run ** on a computer (possibly together with some kind of simulation ** of the world. In that case, of course, the robot's cognition ** becomes indistinguishable from a program's, and my point is made.

[SH: Nothing of the kind. You can't deduce the conclusion you are trying to draw here except by presupposing it, as you do: In the "biological" case it's no-go because a brain is not a digital computer, and it doesn't just run software. In the hypothetical robotic case we don't KNOW what kind of device would pass the TTT but we do know one thing: it couldn't just be a computer crunching symbols (for the reasons repeatedly discussed in connection with the symbol grounding problem, transduction, analog processes, etc.). Now, whatever nonsymbolic function it takes to pass the TTT, all bets are off when you NO LONGER have a system that can pass the TTT (e.g., a "robot-brain" in a vat)! If you think that a pure symbol-cruncher -- the kind Searle has shown could not have a mind even if it could pass the TT -- would have a mind despite the fact that it couldn't pass the TTT and despite Searle's argument (and the symbol grounding problem) then you certainly need a stronger argument than mere repetition of the claim that it is so, despite it all. SH.]

** AMR: Third, we need not assume that the entire behavior of a robot ** passing the TTT is being controlled by the Searle-in-the- ** Chinese-Room. It is entirely possible to assume that all ** the functions involving transduction are handled the normal ** way but a few linguistic functions (even one) is handled by ** Searle. The argument can still be made, in that case.

[SH: Searle is already out of the picture when it comes to robots and the TTT (as opposed to pure symbol crunchers and the TT) because he cannot implement transduction and analog processes the way he can implement symbol crunching. So here the systems reply does work: He's not doing everything the system does. But the system is not just manipulating symbols, so that doesn't help the symbolic hypothesis either! Again, you seem to be missing the logic of the arguments and counterarguments. SH.] ** AMR: Fourth, I think the whole line about simulation of processes ** such as seeing as opposed to cognitive ones is misleading. ** As I have pointed out before, I regard simulation as a word ** referring roughly to imperfect--but useful--replications. ** What makes them useful is that users can abstract away from ** the differences between the simulation and the thing simulated, ** but otherwise they are just bad, incomplete replications. ** The question that we need to ask then is what is needed to ** replicate sight as opposed to replicating thought, and ** I am not at all sure that there is any difference in principle.

[SH: What's at issue is not "replication" in some vague general sense, but a specific kind of simulation, namely, symbolic simulation, as in a computer or Turing Machine, where the syntactic manipulation of symbols on the basis of their shapes alone can be given a systematic semantic interpretation. Now it seems (according to the Turing/Church thesis, etc.) that EVERY (finite, discrete) process can be simulated symbolically (and continuous infinite ones can be approximated arbitrarily closely). None of that is in doubt. What is in doubt is that every symbolic simulation has all the relevant properties of the process that it is simulating. Searle showed that this was not so in the case of thinking, but we need not go so far, because, as I've reminded enthusiasts time and again, symbolic furnaces and symbolic airplanes, even though they are semantically interpretable as heating and flying, don't really heat or fly. By exactly the same token, symbol crunchers that are semantically interpretable as thinking don't really think. So much for replicating thought -- and that's without even getting to sight (in which I argue that thought is grounded). SH.]

** AMR: It is an empirical question what kind of phenomenon it is ** we are dealing with. For example, one might have thought that ** sight involves some molecules of the object seen reaching our ** eyes. Well, that is not so. On the other hand, smell does ** work precisely that way, although one might not realize it. ** On the other hand, in a time long long ago, if asked what it ** means to calculate, someone might have insisted that fingers ** or an abacus or a mind (!) is required. Now we know better, ** but Turing's discovery that calculation does not depend much ** the stuff you are made of was precisely that, an important ** and essentially empirical discovery.

[SH: Yes, but where does it follow that thinking in general is just calculation (i.e., symbol manipulation)? Or even that MENTAL calculation is just calculation? SH.]

** AMR: This is a nice case ** actually because, as I pointed out in some earlier note, ** there are notions of calculation (computation would be a ** better word here) that are not physically realizable at all ** (nondeterminism) or only realizable in a physically very ** special way (true randomness). So in those (perhaps marginal ** and perhaps not so marginal) cases it DOES matter what stuff ** you are made of.

[SH: The issue isn't whether or not mental functions are medium-independent but whether they are purely symbolic. Symbolic functionalism is not the only brand of functionalism. There might be functional features of transduction that transcend the medium and that can be implemented in several functionally equivalent ways -- but none of them will be just symbolic. SH.]

** AMR: There is no reason why we should prejudge ** whether thought is a molecular phenomenon like smell, ** or an energy-based one like sight, or one like computation ** where usually--or even always--it does not matter what you ** are made of. Unfortunately, I don't think anybody has advanced ** any cogent arguments to the effect that only someone put together ** say out of mostly water can think. It is quite clear that people ** who mumble something about biology have no clearly defined position. ** Transduction by itself is not enough. You need to show without ** begging the question that transduction is necessary and that ** simulated transduction is not enough. But, as I pointed out, ** while I can see having qualms about the simulability of a human ** brain, I cannot see that in the case of a robot's program running ** the robot in a real world or in a simulated world on a computer.

[SH: This talk about the medium is a red herring; Searle has repeatedly agreed that the brain and lots of other things might be implementable in radically different ways yet still have the same causal powers. He's just concerned with ONE kind of way that WON'T work, and that's symbolic simulation. (Simulated robot worlds have been discussed repeatedly in this forum. They fail, for Searlean reasons, and because of the symbol grounding problem; however, in principle they could test and anticipate what the functional requirements would be to build a nonsymbolic implementation that WOULD have a mind -- just as they could test and anticipate the properties of successfully implementable furnaces and airplanes without actually heating or flying. SH.]

> AMR: Two, the TTT is much too strong,
> since no one human being can pass it for another, and we would not be
> surprised I think to find an intelligent species of Martians or what
> have you that would, obviously, fail abysmally on the TTT but might pass
> a suitable version of the ordinary Turing Test.

SH: Pass it for another? What does that mean? The TTT requires that a candidate be completely indistinguishable FROM a person, TO a person, in all human robotic and linguistic capacities.

You seem to be misunderstanding the logic and the methodology of the TTT. The TTT applies only to our species, and other candidates from which we cannot be distinguished. It is based on the simple intuition that, whatever it is that convinces us that anyone other than ourselves really has a mind, it can't really be anything we know about how the mind works internally (since we don't know how the mind works internally). Hence if a candidate acts (for a lifetime) indistinguishably from a real person with a mind, like ourselves, then we are in no logical or methodological position to deny that he has a mind because of any arbitrary notion we have about how he works (or should work) internally. What distinguishes the TT from the TTT, however, is that, be it ever so convincing, a lifelong penpal test is not strong enough, and not really the one we use with one another. We ask for indistinguishability in all interactions with the world (including robotic ones).

Other species, including extraterrestrial ones, are simply irrelevant to this, both logically and methodologically. In point of fact, we have trouble Turing Testing other species because we don't have the intuitions about what it's like to BE them that we do have for our own species. Nor can we ever know their ecology well enough to know whether they are Turing indistinguishable from their OWN kind. (For similar reasons I will only bring up if someone asks, congenitally deaf/dumb/blind people, paraplegics, etc. are not relevant to this discussion either.)

** I see that Martians would not be equivalent in behavior to ** human beings and so they would be unlikely to pass the TTT. ** But that is my point exactly: you should not have to be ** indistinguishable from a human being in order to think or ** have other cognitive states. In fact, most of us seem to ** agree that animals feel even if they do not think. Yet ** existing programs AND robots do NOT feel. Likewise, deaf, ** dumb, sleeping, etc., people are intelligent, think, feel, and ** so on, yet they are quite different behaviorally from "normal" ** people. So, again, unless the purpose of the TTT is to identify ** "normal" people, it serves no purpose. But in any case, it does ** NOT serve the purpose of distinguishing thinking from unthinking, ** feeling from unfeeling, intelligent from unintelligent, entities. ** The test is too strong for the purpose at hand.

[SH: The human TTT may well be too strong, but it's the only one we know intuitively how to judge, and anything less is either too weak or simply indeterminate. Besides, even the TTT is not so strong as to amount to a guarantee. Sub-TTT performance, however (including the TT, if it can be passed with just symbols), is much more likely to lead (as it has so often already) to spurious overinterpretations of mindless toy performance. SH.]

> AMR: Third, the TTT is still
> a criterion of equivalence that is based exclusively on I/O, and I keep
> pointing out that that is not the right basis for judging whether two
> systems are equivalent (I won't belabor this last point, because that is
> the main thing that I have to say that is new, and I would hope to address
> it in detail in the near future, assuming there is interest in it out
> there.)

SH: For reasons already discussed in this symbol grounding discussion group, I don't find your arguments for deeper equivalence convincing. There is, of course, a logical possibility that of two TTT-passing robots, one has a mind and the other doesn't. How could we ever know? Because one of them is more brain-like in its inner functions? How do we know how much brain-likeness is necessary or sufficient? And if insufficient, how do we confirm that it did not, indeed, succeed in generating a mind, despite the TTT-power? It seems as if your stronger functional equivalence criterion is an a priori one, and arbitrary, as far as I can see, for there is no way to determine whether or not it is really necessary.

Of course, it's always safer to ask for more rather than less. What I've called the "TTTT" requires indistinguishability not just in robotic performance capacity (what our bodies can do) but in neural performance capacity (what our cells and even our biomolecules can do). That's as much as empiricism and objectivity can furnish. I'm betting, however, that this extra fine-tuning constraint will be unnecessary: That once the problem of generating TTT performance itself is solved (no mean feat, despite how easy it is to suppose it accomplished in thought experiments like Strong AI's and Searle's), ALL the successful implementations will have minds (I call this "robotic functionalism"), and that aiming at generating TTT-capacity will be a strong enough constraint to guide our research.

What’s sure is that we can never know the difference (except in the special teleportable case of symbol crunching). This is one of the reasons I am a methodological epiphenomenalist.

** AMR: The reason you worry about deeper notions of equivalence being ** untestable (unempirical) is I gather the same as Turing’s. He ** also argues quite explicitly that if there is a distinction ** that is not reflected in I/O behavior, then it makes no sense ** to worry about it. However, this is a misconception. Linguists ** routinely talk about strong as opposed to weak equivalence of a model ** with a native speaker (alias ** descriptive vs. observational adequacy), and they seem to be making ** sense. For example, it does make sense to claim that English ** speakers in some sense know an s suffix for plurals or even ** that we interpret children as child + ren and not (as the ** etymology is) as childr + en. A nonlinguistic example might ** be in order: We observe that many programs running on PC’s ** prompt people by saying things like Strike any key... Now, ** we could imagine that the PC keyboard sends a distinct signal ** for each key pressed, but then we would probably not see ** many programs doing this. It would be too cumbersome to program ** and the convention would never have evolved. On the other hand, ** we can imagine (and this is right) that each PC key basically ** sends two signals (one of which is the same for all keys). My ** point is that we can imagine two machines that behave identically, ** yet there can be grounds for deciding that the real machine we ** are looking at is probably like one and unlike the other one. ** How we make such choices is a complex issue, but I would ** focus on the fact that usually we are modeling a population of ** distinct but related entities (e.g., many speakers of one language ** (who are all different), many different languages, the same language ** at many different times in its history) and we form theories that ** are supposed to account for the whole population. Plus, we usually ** try to imagine how the mechanism we are postulating could have ** evolved. Thus, precisely by considering what Martians and human ** beings or deaf and hearing human beings or just Chinese and English ** speakers have in common, I believe that we can choose among ** numerous weakly (I/O) equivalent models that are strongly non- ** equivalent, and still stay empirical.

[SH: There’s nothing wrong with adopting more empirical constraints: We can demand that our models explain not just immediate performance, but history, evolution, individual differences, reaction times and even the effects of a kick to the left side of the head. All those constraints cut down on the degrees of freedom and make it more likely that we’ve captured a real mind. I’m just betting that most of the degrees of freedom will already be pinned down by what it takes to successfully generate TTT power itself and that the rest will just be the fine-tuning. (By the way, just as symbolic simulation can test and anticipate the road to a successful nonsymbolic implementation, so some of the constraints you mention might suggest ways to successfully pass the TTT. But that still leaves the burden on the TTT, rather than the signposts. As a matter of fact, though, very little help along the path of the TTT has come from, for example, neuroscience so far; on the contrary, modelers seem to be giving neuroscientists a better idea of what to look for.) SH.]

> AMR: (3) Likewise, I omitted to refer to Harnad’s position on simulation
> because (a) I thought I could get away with it and (b) because I do not agree
> with that one either. The reason I disagree is that I regard simulation of a
> system X by a system Y as a situation in which system Y is VIEWED by an
> investigator as sufficiently like X with respect to a certain (usually
> very specific and limited) characteristic to be a useful model of X. In
> other words, the simulation is something which in no sense does what
> the original thing does. However, a hypothetical program (like the one
> presupposed by Searle in his Chinese room argument) that uses Chinese

> like a native speaker to engage in a conversation that its interlocutor
> finds meaningful and satisfying would be doing more than simulating the
> linguistic and conversational abilities of a human Chinese speaker; it
> would actually be duplicating these. In addition--and perhaps this
> is even more important--the use of the term simulation with respect to
> an observable, external behavior (I/O behavior again) is one thing,
> its use with reference to nonobservable stuff like thought, feeling,
> or intelligence is quite another. Thus, we know what it would mean
> to duplicate (i.e., simulate to perfection) the use of a human language;
> we do not know what it would mean to duplicate (or even simulate partially)
> something that is not observable like thought or intelligence or feeling.
> That in fact is precisely the open question.

SH: "Viewing-as" is just hermeneutics. It is not a matter of interpretation whether I, as a "model" of someone who understands English, am really understanding English. I either am or I am not. Searle's special little privileged periscope is here to tell us that implemented symbol crunchers do not. That normally unobservable stuff is, after all, observable to ONE observer, namely, the subject -- if it's there, that is. And Searle is in a position to report that, in a symbol cruncher, it's not. (This is the classical point where people jump in with irrelevant non sequiturs about unconscious understanding, etc. etc. I'll reply only if someone -- sigh -- insists on bringing it up.)

The conclusion, by the way, is not just that symbol crunching is NOT understanding, but also that the TT is insufficient, for at least one kind of candidate could hypothetically pass it while demonstrably not understanding. Hence linguistic performance is no more an isolable mental module than any other fragment of our performance capacity (like chess-playing, theorem-proving or block-manipulation) is. Nothing less than the TTT will do; the rest are just underdetermined (and mindless) toys.

** AMR: I agree that the TT is insufficient, although actually it too is ** also much too strong. It is not crucial to our notion of thought ** or intelligence that the language used be the same. What we ** want to know is what the most disparate kinds of minds have in ** common, so that we may find a way of distinguishing minds from ** mindless machinery. A fortiori for TTT. There is to my mind ** very little if any difference between TT and TTT (in principle), ** and it almost seems like an accident that Turing chose the linguistic ** form of the test instead of something else.

[SH: Not much difference between designing something that talks just like a person vs. something that acts just like a person in every respect? Well, since we're dreaming in both cases it's probably safe to say you don't see much difference. I too happen to believe there won't be a difference, but only because I believe it would take TTT-power to pass the TT in the first place, because of the symbol grounding problem. But Turing's "accident" (the party game) sure has led to a lot of confusion... SH.]

> AMR: And, again, it seems to
> me that the relevant issue here is what notion of equivalence we employ.
> In a nutshell, the point is that everybody (incl. Harnad) seems to be
> operating with notions of equivalence that are based on I/O behavior
> even though everybody would, I hope, agree that the phenomenon we
> call intelligence (likewise thought, feeling, consciousness) are NOT
> definable in I/O terms. That is, I am assuming here that "everybody"

> has accepted the implications of Searle's argument at least to the
> extent that IF A PROGRAM BEHAVES LIKE A HUMAN BEING, IT NEED NOT
> FOLLOW THAT IT THINKS, FEELS, ETC., LIKE ONE. Searle, of course,
> goes further (without I think any justification) to contend that IF A
> A PROGRAM BEHAVES LIKE A HUMAN BEING, IT IS NOT POSSIBLE THAT IT
> THINKS, FEELS, ETC., LIKE ONE. The question that no one has been
> able to answer though is, if the two behave the same, in what sense
> are they not equivalent, and that, of course, is where we need to
> insist that we are no longer talking about I/O equivalence. This
> is, of course, where Turing (working in the heyday of behaviorism)
> made his mistake in proposing the Turing Test.

SH: (1) Your "strong equivalence" criterion is at best a metaphysical insurance policy you can never cash in; at worst it's arbitrary. Performance capacity is the only real arbiter.

(2) No one wants to DEFINE mental states in I/O terms. (Who's in any position to do any defining!) Empirical methodology is the only thing at issue here. You can never know that a stone hasn't got a mind, of course. And even Searle can't prove that that computer over there can't understand even though when he does everything it is doing internally, he doesn't understand. He can't even prove he hasn't got TWO minds, one the regular one, and the other a result of multiple personality disorder engendered by the memorization of too many symbols. So "necessity" and "possibility" are not at issue here; that's too strong.

(3) The only thing Turing got wrong, in my view, was the systematic ambiguity at the root of the difference between the TT and the TTT. And Searle has succeeded in showing (by a purely intuitive argument) that by the usual probabilistic canons of empiricism, it is highly unlikely that either he or that computer understands Chinese. A need for some "stronger" form of equivalence has absolutely nothing to do with it.

** AMR: I would agree with you if strong equivalence were not empirically ** testable, but it is, in the sense hinted at above, where we find ** general theoretical grounds (based on looking at a population of ** different but related objects and on thinking about how they ** evolved) for postulating one structure for the given object over ** another (even when the I/O predictions of the theories are the ** same). I am tempted to point to such classic cases as ** Copernican, Ptolemaic, and Keplerian models of the solar system ** which made the same predictions but still were not equivalent in ** a very real sense. I sympathize with the fear that the notion ** of strong equivalence engenders, but that fear is unjustified. ** If anybody is willing to give the idea a chance, then of course ** it will require more support than I have given it here.

[SH: I think I see the source of our disagreement. I am thinking about the state of affairs when we are near a complete theory -- almost at the end of the path to the TTT. We are of course nowhere near there; we hardly know how to generate any performance at all. Now if any of your considerations of "strong equivalence" help aim us or accelerate us along the path, bravo! Let people gather their creative insights wherever they may. But that's not what's at issue; what's at issue is whether, once we're at or near Utopia, "strong equivalence" is of any consequence, and I doubt it; moreover, whatever help it has been along the way will only be help in virtue of the fact that it has produced performance (TTT) dividends; it has no independent weight, except as TTTT-fine-tuning. SH.]

** But I still say that elements of the TTT are useful if not ** indispensible. It really does make sense to use nonlinguistic ** behavior together with the linguistic, but even if there is ** a distinction in principle (which I doubt), it remains to be ** demonstrated in a way that does not beg the question. ** That there are such distinctions to be made can be shown in ** the following example: ** ** Being of Two Minds ** ** Imagine a French person communicating with two separate ** people sitting in a room somewhere and pretending to be ** a single Frenchman. Assume that neither of the two knows ** French well enough to do it, but that between them they do ** (e.g. one is good at comprehension, the other at production ** of French utterances). Between, they use another language ** (perhaps English) to decide what needs to be said to the ** Frenchman. Now, quite clearly the UNDERSTANDING that happens ** is done individually by the two minds, but KNOWING FRENCH ** PERFECTLY is something that the two minds do BETWEEN THEM. ** (There are more complex variations on this theme. The ** basis of the argument is an intuition like Searle's, except ** that this one is easier to stage (in fact, I have once been ** in something approximating this situation). So, it turns ** out essentially that "understanding French" and "understanding ** what s. o. said (in French)" are very different kinds of predic- ** ations. The former clearly can be done collectively, the second ** perhaps not. So, just as I said, there are cases where we can ** make the relevant distinctions. I am not disputing that. I ** am just disputing the contention that the TTT notion ** makes the right distinction. Alexis Manaster-Ramer

[SH: This example seems to me to be irrelevant to the issue at hand, except insofar as it, like Searle, shows an ABSENCE of understanding in the collaborative interlocutor. I don't believe there is a joint "system" there that understands things that neither of the team understands individually, even if it seems that way. Look, there's no limit to the number of tricks that can be played that result in misinterpretation or overinterpretation. So what? Aren't we trying to get it right, rather than milking our misinterpretations? Stevan Harnad.]

> From: FROSTROMS%CPVB.SAINET.MFENET@CCC.NMFECC.GOV
> Subject: simulation/duplication
>
> I have not been keeping up to date on the grounding discussion, so please
> forgive me if this is silly. I am curious as to your response to the
> following hypothetical scenario:
>
> Let's say we could build a device that would COMPLETELY model human
> neurons, and be programmable to take on the characteristics of a
> specific neuron. That is, for all possible inputs, the device would
> give the same outputs as its organic counterpart.
>
> Further, let's say we could examine the connectivity and information
> associated with organic neurons, such that we could take an organic
> brain and duplicate it component by component. Suppose we extend this
> to an arbitrary level, duplicating components in the body on a one-for-one
> basis, including electrochemical, sensory and manipulative aspects.
>
> Would the resulting system be a true thinking machine in the subjective
> sense, retaining its memories? Would it be able to pass the TTT?
>

> Stephen A. Frostrom frostroms%cpvb.sainet.mfenet@nmfecc.arpa

What you say is not silly, but it's completely ambiguous on exactly the points that all the discussion here is focused on: By "completely modeling" neurons do you mean simulate them symbolically in a way that is INTERPRETABLE as doing the exactly same thing that neurons do (with inputs and outputs and components that are likewise merely simulations of neurons' inputs and outputs and components)? Then the answer is: No, such a system would not be thinking even if it could pass the symbolic TT (symbols in, symbols out), because of Searle's Chinese Room Argument; nor could it pass the (robotic) TTT, because it wouldn't even have the sensorimotor transducers; nor would its symbols be grounded.

If you mean a device that DOES have the capacity to pass the TTT, then you can't just mean a symbolic simulation, and this grounding discussion is about what kind of system that would have to be, with what kinds of properties, in order to be grounded.

Simulating neurons and their connections may help you understand them well enough to help you build a device that can pass the TTT, but the implemented device can't be just symbolically simulated neurons, any more than an implemented airplane or furnace can be just built of just symbolically simulated components.

Finally, I doubt that trying to simulate the minutiae of neurons is the right way to go in trying to build a device to pass the TTT. Trying to generate the TTT performance is the right constraint, not trying to mimic the little we know about the physiological implementation. That's just a hunch about research strategy, however, not an argument.

Stevan Harnad

From: harnad (Stevan Harnad) To: "Keith F. Lynch"

You haven't understood the logic of my rebuttals: There is no evidence whatsoever to support (and plenty of evidence to contradict) that memorizing symbols generates two minds within a brain. Hence when Searle points out that he would not be understanding anything if he did so, it is just sci. fi. to reply that "something" in there would be understanding anyway. That something would be understanding in a pure symbol cruncher was the hypothesis that was on trial here. Searle has provided negative evidence. It is not very good reasoning to simply resurrect the very same hypothesis as the defense against that counterevidence. Think about it...

> What evidence [that memorizing symbols does not generate two minds
> within a brain]? Keep in mind that the scale of the experiment is such
> that it could never really be done.
>
> Scale often confounds intuition. Who would guess that (to use the
> Sci.Am. analogy) waving a magnet quickly would produce light? Or
> that a bunch of people together in the wilderness would create a
> technological civilization? Or that a bunch of hydrogen gas in one
> place would become a star? Or that a bunch of hollow tubes with a
> heated cathode, a grid, and a plate (or the equivalent of these tubes)
> could constitute this computer network? Or that a bunch of cells in

> a skull could be a brain which could support a conscious intelligent
> mind?

I know the Churchlands argument, and its no good:

(1) Slow EM oscillation IS "light," it's just not in the visible range. For physics, the fact that you can SEE some ranges is irrelevant. Their example just mixes up the objective physical phenomenon with the subjective mental one.

(2) Although there do indeed exist physical continua along which unexpected phase transitions lawfully occur, these are inferred from evidence, not simply posited as a possibility or analogy in the face of counterevidence.

(3) We can certainly try an experiment memorizing a few symbols, and then more and more. No evidence for a phase transition will occur, and I don't think you doubt it; nor will there be evidence for subthreshold understanding, as in the EM spectrum. The only basis for supposing understanding will be sci fi fantasy, based on analogy with overinterpreted "virtual" systems in computers and irrelevant emergent phenomena in physics. Counterfactual conjectures are not the right logical or empirical response to counterevidence against a hypothesis. The right response is to find stronger positive evidence, revise the theory or reject the hypothesis.

> What counterevidence? All I see is assertion and counterassertion.
> Neither side has proven anything. We simply differ over which is the
> null hypothesis, and on whom is the burden of proof.
>
> If a pure symbol system can't have semantics, what sort of physical
> system can, and why?

If Searle's thought experiment was not convincing enough, get subjects to start memorizing symbol manipulation rules and see if you start getting either subjective testimony about emergent understanding or clinical evidence of an emergent cognitive dissociation (split personality). If you think a phase transition would occur beyond your experimental range, present evidence in support of THAT.

I've already answered the question "If a pure symbol system can't have semantics, what sort of physical system can, and why?": A hybrid symbolic/nonsymbolic system grounded in iconic projections and category invariants. And because such a system could actually pick out the objects and states of affairs to which its symbols refer (and because such a system would be immune to Searle's argument, as my transducer counterargument shows).

Stevan Harnad

> From: ganesh@cs.wisc.edu (Ganesh Mani)
>
> Stevan - Here are some of my thoughts/arguments on the Chinese Room
> Argument and on characterizing ‘‘intelligence’’. Let me know what you
> think of them. I am in the process of expanding them and presenting
> them in a more coherent form. Hence your criticisms and suggestions

> would be appreciated. I guess you could post what I sent you to the
> symbol grounding group.
>
> Thanks -Ganesh Mani
>
> The Chinese Room Argument fascinates, bewilders and elicits a shrug
> from (implying that the argument is not particularly germane) people
> involved in the field of Artificial Intelligence.
>
> Of course, I (like a number of others) have problems with Searle
> completely ignoring the complexity issue. Human beings are adept at
> finding reasonable solutions to everyday problems such as packing ones
> suitcase, which a theoretical analysis would show to be hard
> (NP-complete). One aspect of intelligence (whatever this elusive entity
> might eventually turn out to be), thus involves efficiency. For
> example, it would be completely unacceptable if the ‘‘Turing Test
> candidate’’ took 50 years (or some such ridiculous amount of time) to
> respond to each question. I would not pass the candidate with
> satisfactory answers to one or two questions and if the candidate
> answers more than two at that rate, there is all the more reason for
> not passing it/him/her.

Irrelevant. The conjecture is that a pure symbol cruncher can pass the Chinese TT; Searle accepts the conjecture for the sake of argument (it may be counterfactual -- I think it is) and then proposes a thought experiment in which he does the same thing the symbol cruncher does. It is easy to see that he would not be understanding Chinese under those circumstances. The fact that he may not actually be able to crunch symbols fast enough in reality is completely irrelevant. There’s nothing in all of this that warrants hypothesizing some kind of "phase tranasition" (into mental space?) with speed -- or complexity, for that matter. That’s just hand-waving, and theory saving.

> However, I digress; to get back to Searle, the two approaches I would
> take to force Searle to understand or to prove that he would understand
> if he were passing the Turing Test in Chinese are:
>
> 1) Assume that I am interrogating Searle in Chinese by writing each
> question on a separate card. Give Searle Chinese symbols (on a separate
> card before I have received my response to the previous card) saying to
> the effect ‘‘Stop working on the previous question and don’t give a
> response to either the previous question or this one’’. Assume the
> English speaking Searle would not be arrogant or persistent to keep
> replying to the ‘‘Stop’’ card. (Can this be bolstered by the threat
> that he would actually not pass the test (TT) if he kept doing
> that---leads into the conditioning argument). If so, the Chinese Searle
> should also not reply to the ‘‘Stop’’ card. Now Searle knows the
> meaning or somehow has associated the ‘‘Stop’’ card with the act of
> curtailing his response. Can we hence say that Searle knows the meaning
> of one Chinese word, ideogram or sentence? Using this, it can be argued
> that Searle can bootstrap his way up exactly like he talks about being

> at different levels (as far as understanding goes) with respect to
> German and French. Is this similar to the TTT since Searle has to
> perform an action (actually in our case, refrain from performing an
> action) after seeing the ‘‘Stop’’ card ?

Irrelevant again. The question was not whether Searle could eventually learn Chinese in the Chinese Room. The question was whether it is true that to do the symbol manipulations -- to implement the right program, just as the computer that passes the TT also does -- is to understand. Searle shows it definitely isn’t. What Searle may or may not go on to learn under these conditions is irrelevant; the claim wasn’t that the computer would learn to understand, it was that the right symbol crunching IS understanding. (If the claim HAD been that the computer eventually learns to understand, that would be wrong too, by a simple recursive application of Searle’s Argument.)

The problem of learning a second language from symbols alone ("the Chinese/Chinese Dictionary-Go-Round") is given as an example of the symbol grounding problem in the article under discussion in this group; it is soluble, as prior feats of human cryptology have shown; but its harder variant -- learning a FIRST language from symbols alone is insoluble. That’s the problem facing the pure symbol cruncher.

> 2) To condition Searle using, say, a bell (the sound of a bell can be
> thought of as a symbol being communicated through a different channel,
> the auditory channel). Slowly Searle would respond with a set of
> symbols to the bell alone. (or would he?) --- If he didn’t, he would
> reduce his probability of passing the TT since I know that a
> conditioned response can be elicited from most human beings. (This
> argument needs to be strengthened but the point I am trying to get at
> is to imbue understanding in the system via conditioning)

There is no argument here; and the TT (the purely linguistic, "pen-pal" version of the Turing Test -- symbols in, symbols out) has nothing to do with Pavlovian conditioning. The TTT -- the full robotic version, the Total Turing Test -- would indeed have to be able to display Pavlovian conditioning, but we already know that that that test is (a) immune to Searle’s Argument (because of the Transducer Counterargument in my "Minds, Machines and Searle.") and (b) could not be passed by a pure symbol cruncher in the first place.

> The idea in both 1) and 2) is to make Searle associatively learn the
> meanings of (or understand) a few concepts and then try to bootstrap on
> them in an effort to inculcate in Searle (or any system that is being
> trained) the ability to understand more and more (or enough to pass a
> sufficient performance test of intelligence).

To repeat: Strong AI or the Symbolic Theory of Thinking was not about training systems but about thinking being symbol crunching. The theory is wrong; training is irrelevant, and would not help save the hypothesis. Moreover, if the learning system was not a pure symbol cruncher, it wouldn’t even TEST the hypothesis.

> Specifics:
>
> Assertions

>
> 1. A formal system can specify how semantics are to be attached to the
> formal (syntactic) elements.
>
> 2. Thus, an extensible formal system can have meaning, semantics or content.
>
> 3. If the content is the same as the meanings that run through a human
> mind, they share a frame of reference.

These are just bald assertions. You must consider the arguments that have already been adduced against them in this discussion, repeatedly, viz:

(1) A symbol system is ungrounded; it specifies nothing except with the aid of an interpretation, and it is the interpretation that is ungrounded.

(2) No purely symbolic "extension" of a symbol system is grounded either.

(3) There is no "content" to a symbol system other than the interpretation that is projected on it by human minds. This is the hermeneutic hall of mirrors. The "content" is only in the mind of the beholder. ("Frame of reference" is merely a figure of speech, so far as I can tell.)

> Conclusion
>
> Minds and machines can be indistinguishable (as the Turing Test
> implies) with respect to this “frame of reference” and, hence, should
> be deemed equally intelligent.

Bald assertion again, not taking into account the counterevidence, viz., Searle’s Chinese Room Argument and the Symbol Gorunding Problem.

> Searle further avers that “Any function that can be computed on a
> parallel machine, can also be computed on a serial machine”. This is
> certainly true in almost all instances; however, the effect of two
> computations (one serial and the other parallel), taken in the context
> of the environment the computations are embedded in, can be different.
> For example, if the results of the computations are going to be
> interpreted at different time instants (which will have to be the case
> since the serial machine will usually be slower), they will have
> different interpretations since the world or the frame of reference is
> different at different time instants.

Irrelevant; speed is not at issue; the grounding of the interpretation is. If parallelism (or speed) are critical to implemeting a mind, this requires some strong, specific argument or evidence. There is none; it’s just a theory-saving measure.

> Computationally, Searle stresses, serial and parallel systems are
> equivalent. As pointed out earlier, different interpretations may be
> attached to the same formal computation as the interpretation is
> usually a function of both the computation and the situation, frame of
> reference, or environment.

>
> Consider the computation ‘‘Is this computation sequentially
> implemented?’’ This will result in a ‘‘yes’’ from the serial
> implementation of the computation and a ‘‘no’’ from the parallel
> implementation. Thus, self-referential computations may not satisfy
> Searle’s assertions.

Irrelevent. It is not part of the TT or the TTT to explain correctly how your mind works otherwise none of us would pass.

> On Intelligence:
>
> a) Intelligence has the property that it is situated. It has to be
> measured with respect to a frame of reference or domain of interest.
> This implies that in a purely symbolic world (such as the world of
> certain IQ tests which rely only on symbol processing such as pattern
> matching, extension, and induction), intelligence can stem from pure
> symbol manipulation alone. If the performance domain is the toy blocks
> world domain, an agent to stake its claim for intelligence, will have
> to successfully complete certain tasks such as perceiving, moving and
> stacking the blocks as may be required by the goals of the performance
> task. In the world we live in, intelligence may be thus attributed to
> agents performing a certain number of tasks effectively from a set of
> tasks human being performs.

This is all metaphor. There’s only one world, but we have several tests. The TT tests only our pen-pal capacities, the TTT our full robotic ones. The TT could conceivably be passed by symbol crunching alone, but then, as Searle has shown, the passer would not have a mind. In this sense the TT is not decisive. The TTT is no more decisive, but it’s the best we can expect (other than the TTTT, which asks for neuromolecular indistinguishability, but that wouldn’t be decisive either, because of the "other minds" problem). So, until further notice, only minds are "intelligent," and only devices that can pass the TTT have minds. The only relevant "situatedness" is the grounding of a robot and its symbols in the world (the only one there is). "Frames of reference" and "domains of interest" are just in the minds of the interpreter if there is no TTT-scale grounding.

> b) A computational model of intelligence should also capture the time
> aspect, penalizing slow reasoning when the domain demands fast
> reasoning.

Yes, performance timing is part of the TTT, but that’s irrelevant to any of the other issues at hand.

> Meta-level comments:
>
> The extensible formal system I am talking about might well map into
> your grounded symbol system.

I’m afraid you haven’t understood the grounding problem or my proposal at all if you imagine that a pure symbol system can have semantics.

Stevan Harnad

DIRECT VS INDIRECT GROUNDING

> From: lammens@cs.Buffalo.EDU (Joe Lammens)
>
> Perhaps this has been discussed before, in which case I apologize. I
> haven't followed the discussion from the beginning.
>
> I was reading Jackendoff's "On Beyond Zebra: The relation of
> linguistic and visual information" (Cognition 26 (1987) 89-114). He
> essentially advocates coupling Marr's 3D vision models to his own
> semantic/conceptual structures for language, through a translation
> process. One of the advantages of doing so would be that "semantics
> will not have to back up into ever murkier levels of 'interpretation'
> to explain our ability to talk about the visually perceived world" (p.
> 93).
>
> Now this seems to be a proposal for grounding semantic/conceptual
> symbols in (models of) the visual, i.e. perceptual, world. Do I
> understand correctly that that is what you would take it to be?
>
> It's a very interesting proposal indeed, but it does not solve all
> problems in terms of symbol grounding. There are obviously abstract
> concepts for which there can be no (direct) grounding in (models of)
> visual perception, and it is not clear how abstract properties like
> color would be grounded in this way, although the property is clearly
> related to (derives from) visual perception. I hasten to add that
> Jackendoff does not claim to be able to "ground" everything in the
> semantic/perceptual domain through a translation into/from the 3D
> models domain.
>
> The question then: I would like to know your thoughts on how to ground
> symbols representing abstract concepts, properties and the like in
> perception and/or action. Is there some sort of hierarchy with
> grounded symbols at the bottom and symbols at higher levels built out
> of them, being "indirectly grounded"? There is some neurological
> evidence for hierarchical structuring of (visual) perception, for
> instance in the work of Hubel and Wiesel. But there is no agreement in
> the field on its significance or validity I understand.
>
> Joe Lammens.

In the paper under discussion, "The Symbol Grounding Problem" I discuss the "zebra hierarchy": Suppose the symbols for "horse" and "stripes" are elementary and grounded (i.e., one has iconic representations for discriminating them and invariant categorical representations for identifying them from their sensory projections). Then one gets the new symbol "zebra" for free, just from the symbol-string: "Zebra" = "horse" & "stripes." My claim is that you can get to the abstract concepts of goodness, truth and beauty exactly the same way. -- Stevan Harnad

----

> From:
> Subject: sym grounding talk at iu
>
> I wanted to thank you for the talk you gave the other day here.
> I've read two of your recent papers (JETAI 89 & "The Symbol gnding prb")
> and have a few comments. Also, I got a bootleg copy of correspondence
> that you were redistributing over the Internet on the subject--I
> wanted to ask if you could add my name to your DL for chinese room
> discussion? The last message I saw was for December or so...
>
> Anyway, I'm inclined to agree with you that something like
> your symbol grounding mechanism is the (or a) right way to
> introduce semantics into the symbol system. In fact, it sounds
> a lot like what philosophers in general would have in mind when
> they start talking about a "causal theory of reference"--although
> worked out in some modest detail (i.e., infinitely more than
> philosophers are used to). Does your approach have specific
> antecedents in philosophical theories of reference? Could you
> give me references so I could follow the trail sometime? (& thanks)
> (and I understand that I should but haven't yet looked at your
> book Cat Rep).

It's a form of causal theory of reference, but a rather specific one: The relevant causal property is the capability of discriminating and identifying (categorizing) the objects and states of affairs that are referred to. At "ground" level this is done on the basis of the sensory projection, but then the abstraction hierarchy can be scaled through by concatenating (grounded) symbols so as to define new category memberships, no longer directly sensory ones.

I'm always being told that this or that is a philosophical antecedent (usually one that has allegedly been shown to be inadequate) -- empiricism, verificationism, procedural semantics, etc. etc. -- but usually the similarity turns out to break down or to become trivial. It's certainly not verificationism, for example, because for me a "peekaboo unicorn" -- defined as a unicorn that disappears tracelessly whenever any sense organ or measuring instrument is directed at it -- is a perfectly well-grounded concept, yet necessarily unverifiable.

> Transduction & analog devices: at the meeting Dave Chalmers went on
> how about how you required an analogue subsystem to process the
> distal projections into icons, etc., while you denied that and
> said you're only committed to non-symbolic transduction. I like
> this latter position of yours, but I can understand how Dave
> thought you were holding a different one. In fact, in the
> symbol-grnding paper (p. 17) you say: "In a pure symbolic model
> the crucial connection between the symbols and their referents
> is missing..." Perhaps you would like to say in a PURE symbolic

> system there are no transducers. But I would say in a PURE
> symbolic system there aren't any IO devices of any kind. So
> who cares about purity such as that? It has to be wrong (i.e.,
> a straw-man argument) to saddle traditional AI with the task
> of creating an intentional system without any possible interaction
> with the world. Anyways, in your paper you go on to talk a lot
> about the role a connectionist net might have in perceptual proc-
> essing. I don't disagree with that, so long as you acknowledge
> as well that (in theory at least) non-symbolic digital processing
> can play the same role--not just in simulating a neural net, but
> in doing bit-mapped graphics, for example.

First of all, a transducer is one form of analog system; so it's not a question of transducers OR analogs. Second, of course an implemented symbol cruncher like a digital computer has transducers -- it's a physical system, after all, not an abstract, immaterial one -- but they're the wrong kind of transducers, and they're doing the wrong kind of thing. It's the transducers/effectors of a (TTT-passing) robot that are relevant here. And I'm not sure what you mean by "non-symbolic digital processing." A system either is an implementation of a systematically interpretable symbol system or it isn't. If it is, its hardware is irrelevant and you have a symbol cruncher; if it's anything else, you don't. The status of connectionism is as yet unclear. Symbolically simulated nets are clearly symbol systems, but other implementations of them (say, with real parallelism) may not be. If properties of the nonsymbolic implementation are causally relevant to what they can do then the parallel implementations must be able to do things the symbolic ones cannot.

> This latter point bears on the claim you made that all (anti)
> responses to Searle come down to the Systems Reply. I disagree
> completely. The systems reply dominates the thinking of alot
> of ai types, but the Combination Robot/Systems reply brings
> in something which is new and, I claim, fundamentally different:
> causal connection to the external world. Since the whole
> system includes (potentially) sensory projection, icon,
> etc. (whether handled by analog or digital processing), and these
> are causally related to the external object, Searle can no longer
> claim--as he does to the simple robot reply-- that the causal
> connection plays no role. It clearly does for the system; as you
> say, he (if still in the room) is not doing everything the
> system is doing. If you know of something wrong with my claims,
> I would certainly like to know what....

I couldn't agree more; in fact that's what I wrote in Minds, Machines and Searle. So what's the disagreement?

> &: zebra = horse & stripes. van Gelder wanted to know how '&'
> is grounded. But surely not every symbol in the symbol system
> needs to be ground!? Some can be implicitly defined by their
> functional role within the grammar. Connectives being obvious
> candidates for that. Of course, there's the additional
> question of where does grammar come from; but that's as much
> a question for anyone else as well.

Agreed

> I'm slowly trying to come to terms with van Gelder here. Maybe
> someday. He seems to think that if you shake syntactic tokens
> hard enough, they'll generate some semantics in the infrared.
> But I don't understand what that means.... Regards, Kevin Korb

Neither do I.

Stevan Harnad

> From: Drew McDermott
>
> Arguing with you is exasperating because I have to keep going back
> over points I thought were settled. E.g.,
>
> The hypothesis was T[hinking] = S[ymbolic] C[omputation], and as
> long as we considered only clever computers that couldn't tell us
> otherwise, there was evidence to support it. Then Searle pointed
> out that, according to this hypothesis, he ought to understand Chinese
> too if he memorized a bunch of meaningless symbols. The original
> hypothesis had not anticipated that; in fact, as long as the symbols
> were on the black-board, the "systems repliers" agreed that Searle wouldn't
> understand, but that Searle plus the blackboard would (the other-minds
> problem of course leaves room for that). But then when Searle pointed
> out that he could memorize the symbols, they replied that, well then,
> he would have TWO minds.
>
> This is all wrong. No computationalist ever agreed with the
> prediction that a person executing an understanding program would
> report understanding Chinese. The prediction has always been that
> executing the appropriate program will bring a person into being, and
> that "virtual person" would understand Chinese. It's quite misleading
> to say "he [Searle] would have two minds." It doesn't mislead *me*,
> but it obviously misleads you.
>
> You see why I'm tempted to stop right here and see if we agree on what
> we're talking about before I go on. I think we've really arrived at
> the nub of the dispute: exactly what prediction computationalism would
> make in the CR situation. If you agree with my formulation of the
> prediction, then I can go on to address some of your objections; or
> perhaps you'd like to rephrase them. -- Drew

I do think some systems repliers have claimed the blackboard plus Searle was the system, but I agree that all claim that it would be whatever system implemented the symbol crunching that would have (be?) a mind. So I agree to that; but now, how does one go about testing it? Clever performance is not enough to confirm or disconfirm this hypothesis: How are we to check whether it's CORRECT? If Searle's periscope is not good enough, what is? I mean, when we adopt a hypothesis there's more to it than simply supposing it to be true. How would one go about showing this one to be false? (And don't reply that as long as it can do everything that's smart, it's confirmed; it could just as well be doing the smart things mindlessly; besides, the question is what would DISconfirm it?) I think Searle got the answer right: If you could BE the system yet not have the mind, that would disconfirm it; and that's what his thought experiment shows. What do the systems-repliers answer: "Well, he's being the wrong system then, because EX HYPOTHESI, the right system DOES have a mind." Well there's sure no way out of that one. It's what I've dubbed the hermeneutic hall of mirrors.

And to prove that SOME systems repliers at least feel the tug of what I'm alluding to, don't forget that some have resorted to replying that, for all we know, if Searle memorized enough symbols and did it all fast enough, he WOULD understand Chinese (and I'm talking about the regular Searle now, not some hypothetical systemic alter ego).

Stevan Harnad

> From: sticklen@cpswh.cps.msu.edu (Jon Sticklen)
>
> I think you are making a common mistake when you try to make the
> analogy between some model of a physical prosess and the process itself
> (like a model of a star versus a REAL star) and some model of cognition
> versus the REAL thing.
>
> The problem is that it is very easy to tell the difference between a
> REAL star and a model of a star. for one thing the REAL star makes heat
> and light, whereas the model of a star produces (probably) computer
> output about heat and light.
>
> But tell me what is the objective difference between a (hypothetical)
> working model of cognition and the REAL thing? the output of both are
> in the same terms. the intermidiate results of both are in the same
> terms. if you want to define the REAL thing as precisely what exisits
> between our ears, then of course the model of the REAL thing will not
> make. but other than that way, which i'm sure you don't mean, can you
> give me a clear, simple statement of the difference between the REAL
> thing and the computer model? ---jon---

The difference (for groundedness) is the difference between (1) the robotic TTT and (2) the symbolic TT or a symbolic simulation of the TTT: (1) can pick out the objects and states of affairs in the world that its symbols refer to, (2) cannot. -- Stevan Harnad

----

> From: "Wlodek Zadrozny"
> Subject: Goodman's "grue" paradox -- is it relevant to grounding?
>

> I didn’t see any reference to Goodman in the discussion on grounding.
> But his paradox can be solved by an appeal to "ostension", cf. e.g.
> Rescher "Peirce’s Philosophy of Science", which seems to be a
> synonym for "grounding".
>
> Appendix. (based on Rescher op.cit.)
> The paradox:
> Let "grue" mean: green before year 2111 and blue thereafter.
> Let "bleen" mean: blue before 2111 and green thereafter.
>
> If we do inductive reasoning on the basis of this color taxonomy,
> we can prove that after 2111 all emeralds will be blue.
>
> The paradox cannot be refuted by forbidding reference to time (2111),
> since from the grue/bleen standpoint it is the standard taxonomy which
> is time dependent.

From the standpoint of a poor psychologist, trying to explain how people manage to discriminate, categorize, name and describe things as they do, the "grue" paradoxes are not very helpful, except as yet another example of the fact that our categories must be provisional and approximate, relative to the sample of instances -- and feedback from the consequences of MIScategorizing them -- that we have encountered TO DATE (which means time is an essential part of the category learning paradigm itself). Stevan Harnad

----

Replies to Chalmers, McDermott and Zadrozny

> From: David Chalmers
> Subject: Intentionality, or Consciousness?
>
> Just a minor but important methodological point this time around.
>
> Oded wrote:
>
> >> The distinction, as one might say, is that "real" intentionality plays
> >> a causal role in the production of behavior.
>
> You [SH] replied:
>
> >No, at least that certainly wouldn’t be my distinction. The difference
> >between really believing/desiring/meaning that X and just being
> >interpretable as-if believing (etc) that X is that there’s a subjective state
> >corresponding to believing that X. If there’s nobody home, there’s no
> >subjective state, and hence only as-if intentionality, regardless of
> >causal role in behavior.
>
> This, I think, is a real source of confusion, and perhaps a tactical error
> on your and Searle’s part. This may be what *you* mean by "real

> intentionality", but it's not what anybody else means. Intentionality, very
> roughly, is all about having some natural referential link between a
> physical/functional state and some state of affairs in the world.
> Intentionality is *not* subjectivity. It's unclear that the mere accompanying
> of a physical state by some phenomenology makes it more "intentional".

You're certainly right that I am in a minority of one on some of these points, but I take that to confirm that, for better or worse, I have an original suggestion to make here: I DO want to argue that the ONLY thing that makes a physical state intentional -- i.e., really meaning or thinking or wanting that that-X, rather than merely being interpretable as-if meaning/thinking/wanting that-X -- is that it is accompanied by the right subjective phenomenology. To claim this is to be prepared to argue that there is no such thing as real intentionality without consciousness -- no such thing as the "right" referential link if there's nobody home.

That's why I reject the idea that intentionality is just internal states with the right causal connections to the outside -- unless the "right causal connection" includes the generation of the right subjective state. I have never, however, claimed that intrinsic intentionality and groundedness are necessarily the same thing. Groundedness is also a matter of causal connection between symbol and referent, but only the robotic connection: the capacity to discriminate, categorize, manipulate, name and describe the objects and states of affairs in the world to which the symbols refer on the basis of their projections on the robot's sensory surfaces. That is indeed a behavioral capacity and a causal connection, but one that is only necessary, but perhaps not sufficient for intentionality, which requires the right subjective state.

To put it another way, if there were no such thing as subjectivity, there would be no problem of intentionality. Intentionality is only the "mark of the mental" in virtue of its special relation to subjectivity: There is something it's LIKE to believe, mean, want, think that X. Real intentional states are like that, as-if ones are not. Without the subjective aspect, there's no basis for the real vs. as-if distinction -- or, more accurately, it's just an as-if distinction rather than a real one. (Think about it before dismissing this view as hopelessly nonstandard...)

> One of the major lessons of the last twenty years in philosophy of mind - some
> might say *the* major lesson - is that the phenomenon of intentionality can be
> separated from the phenomenon of consciousness. So intentionality
> (propositional attitudes, narrow/wide content, causal theories of reference)
> has been studied to death, with consciousness rarely getting a look in. Now
> some, myself included, think that this is a great pity, but what's done is
> done. Intentionality, almost by definition, has nothing to do with
> consciousness.

Whatever spurious independence the notion of intentionality has won from consciousness has been won in virtue of kicking away the ladder that got you to the mental heights in the first place. And what you're left with is hanging by a skyhook. This illusury autonomy is not a pity, it's incoherent, in my view. (Which is not to say that consciousness can be studied: I'm a methodological epiphenomenalist and think that only robotic performance capacities -- TTT, and neural ones, TTTT, if necessary -- can and hence should be studied. The rest must be taken on faith. But the reality of conscious states is not to be denied. The methodological partitioning is a better research strategy than forced (and empty) attempts to "identify" them with something else, or to "project" them onto something else by mere interpretation.)

> Now, you could try to play ideological games, and *redefine* "real, intrinsic
> intentionality" to mean "having a subjective state", or to at least require
> that as a precondition, but it just seems to confuse the issue (e.g., you get
> all these people saying "understanding just *consists in* certain behaviour",
> or "intentionality is merely a certain causal relationship). Instead, why
> not embrace the terms for what you're *really* talking about: "consciousness";
> "qualia"; "subjectivity"; "phenomenology". There's nothing wrong with those
> words, they're perfectly good, respectable if ill-understood phenomena.
> Phrasing everything in terms of "intentionality", "symbol grounding", even
> "understanding" just distracts from the real issue, and leads to endless
> pointless arguments. Leave intentionality be: talk about consciousness.

I'm not embarassed to speak of subjective states. There are sensory subjective states, like what it's like to see blue, affective subjective states, like what it's like to feel angry, and intentional subjective states, like what it's like to understand that all men are mortals. I think the confusion is on the part of those who would hold that the latter is an independent case. Moreover, it isn't as if these wide/narrow functionalisms and causal theories of reference have made much progress in fleshing out intentionality...

And let's not confuse symbol grounding with any of these things. Symbol grounding is a problem for people who want to build robots that will pass the TTT even if they don't know or care whether they will have real intentionality (or any mental state) once they can do so. I just happen to believe they will, and that we are such robots.

> After all, consciousness is what you always come down to, at the bottom line.
> I quote: "really meaning ...that X [means] that there's a subjective state".
> It's unclear that the notion of "intentionality" is doing any work at all.
> How does a subjective state help fix reference? Or help ground symbols, for
> that matter? I don't think there's any *reference* in my experiential
> states, just qualia.

I don't think subjective states have any independent causal power at all. I am a methodological epiphenomenologist. However, I am prepared to trust that they piggy-back on TTT-power for some reason, and that they are what we mean when we distinguish real from as-if intentionality.

> Everybody knows that *really*, the Chinese Room is a problem about whether
> machines could be *conscious* (have qualia, have subjective states...).
> Putting things in terms of "understanding" and "intentionality" bothers me,
> partly because of the consequent confusion, and partly because it is
> reminiscent of the systematic downplaying of consciousness in recent
> philosophy of mind ("ssshhh... don't mention 'consciousness' unless we really
> have to!").

I think I basically agree with you here.

> Side note: if you're looking for more speakers for the SPP Searle session
> (Bob Port told me you might be), I'd be happy to help out. I've done an
> extensive survey of the literature on the Chinese room (over 30 articles,
> not counting those in BBS), and drawn a few conclusions which I've put

> together into a nice little talk called "How Not to Answer Searle". It goes
> through some of the many point-missing booby traps it's possible to fall into
> (we see them on this list all the time), and tries to isolate the real issues.
> And after the negativity, it ends positively: the only *real* way to answer
> Searle is to embrace the Systems Reply for all it's worth, and embrace the
> consequences (such as "multiple people in one head"). I'd count the
> Embarrassment Factor at this counter-intuitive answer as the major reason the
> debate has gone on so long -- almost *none* of those 30+ papers go over the
> multiple personality issue, but this is the only real issue that matters.
> Almost none of the 30+ anti-Searle papers really engage his arguments: the
> only way to do that is provide positive arguments for "two phenomenologies".
> Which I have been working on, and have a very nice little argument worked out.
> Anyway, this will all fit together into a nice short talk, if you're
> interested. >
> I enjoyed chatting with you a few weeks ago. Dave Chalmers.

"Embarrassment" factor is putting it mildly! If you want to show that memorizing symbols can cause multiple personality you've got your work cut out for you. (Unfortunately the SPP Symposium Searle Symposium is oversubscribed. But why don't you try out your arguments and evidence for multiple personality on me and the Symbol Grounding Group. To me, it has almost as little force as countering a proof that a certain assumption implies that 1 = 2 or is as improbable as a monkey typing having typed Shakespeare by asserting "Well then, 1 IS equal to 2 and a monkey DID type Shakespeare." That's called saving an assumption at an inflationary counterfactual price. It's also reminiscent of epicycles...)

Stevan Harnad

> From: "Wlodek Zadrozny"
>
> Let me thank again for the talk. Ben Grosof, Alexis and I just finished
> a long discussion about the argument of Searle's. Here is mine and
> Ben's analysis of it.
>
> 1. One assumes that "mentality" is an attribute of people, but
> it is not an attribute of clocks.
>
> 2. Since an ungrounded program is like a clock (we reason by analogy
> here)--it is a mechanism, even if an abstract one, therefore it can't
> have mental states.
>
> This sounds convincing, but it is not a logical proof. Moreover it is
> possible to make the counterargument that

It's no proof, and it's no argument, and nobody I know has made it. A clock may or may not have a mind. I doubt it, but who knows? A computer running a program may or may not have a mind, who knows? (for the same reason). But a computer running a program does not have a mind (understand Chinese) MERELY BECAUSE IT IS THE IMPLEMENTATION OF THE RIGHT

PROGRAM, because Searle can implement it in the Chinese Room and yet still not understand Chinese. THAT's the argument.

Groundedness is a more specific matter, having to do with the connection between a symbol and the object it refers to, and a robot's capacity to discriminate and identify it from its sensory projection, and how to give it that capacity.

> 3. BUT, analogies can be made only up to a point: although a normal, ie
> small mechanical clock doesn't have mental states, a big one,
> consisting of zillions of parts may. We will use here an analogy with
> the grain and the heap--one grain is not a heap but a zillion is. (Or
> use any analogy in which you have a change of quality when the quantity
> changes).
>
> Therefore, formally, Searle's arguments against strong AI are fallacious.
> The fact that Searle cannot imagine himself understanding Chinese rests
> on the clock analogy, but the heap analogy could help his imagination.
>
> This is not an argument against grounding. I believe grounding is
> psychologically plausible, computationally efficient and can be
> a reason for doing some interesting mathematics. Best regards, Wlodek

Your argument from "quantity" has been made many times before in this discussion group -- about speed, about capacity, about complexity -- and it is equally fallacious every time. It is tantamount to claiming that there is some sort of a phase transition when you get a large enough amount of some nonmental structure or function so that it somehow goes off into mental hyperspace. Perhaps it does. But until some evidence is provided for the existence of such a phase transtition, there is absolutely no reason for believing in it. This is not an argument by analogy but an argument by fantasy. (And you're right, it has nothing at all to do with grounding.)

Stevan Harnad

> From: Drew McDermott
>
> > SH: If Searle's periscope is not good enough, what is? I mean, when we
> > adopt a hypothesis there's more to it than simply supposing it to be
> > true. How would one go about showing this one to be false?
>
> As I've said before, I haven't got the foggiest idea how we would in
> the future confirm or disconfirm conjectures about whether particular
> systems had minds. We're reaching far into the future here, imagining
> that AI has succeeded, and then asking for details on exactly how it
> succeeded. I presume that if we know how to build programs that
> implement minds, we'll also know the difference between real and
> apparent minds. You may think this a presumptuous assertion of faith,
> but it isn't, in the context of a thought experiment, created by
> Searle, in which we imagine exactly this sort of future.

I don't think you're looking closely enough at the hypothesis (or assumption) you've adopted. According to it, having a mind just depends on imlementing the right program. To believe that is to believe that there exists such a proogram. Searle has done no more than to accept this assumption and show exactly where it leads.

> > SH: [and] what would DISconfirm it?)
>
> Again, I can't imagine. It's possible that the supersuccessful AI
> that Searle so generously postulates will predict the possibility of
> pseudominds that can do smart things but give themselves away somehow
> as not really being minds.

I can't follow this: Smart but give-away just means not passing the TT, which means we reject the program in favor of another. The question is how to test and falsify the hypothesis that a TT-passing program has a mind, if Searle's Periscope is not convincing enough. This is, after all, a hopethiesis, ont a definition, isn't it?

> "To be" is not an action verb. I can't "be" the mind in question;
> Searle can't "be" it either. If the theory is true, then
> there will be a virtual person, and the only person that can be that
> person is that person.

In general one can't be the candidate, but one CAN in the special case of Searle's periscope and the hypothesis that every implementation of (the right) program must have a mind.

Stevan Harnad

> From: BARRY MCMULLIN <75008378%dcu.ie@pucc>
>
> I have just got around to a careful reading of the preprint you
> kindly send me ("The Symbol Grounding Problem", Physica D, 1990).
>
> 1: Searle and his confounded Chinese Room.
>
> You find Searle's argument convincing; I find it at best fatally
> flawed, at worst incoherent. My reply to Searle is essentially
> the "systems" reply (to the limited extent that I think a reply
> is possible); but I think that you've had this discussion
> many times already, and its unlikely that a reformulation by me
> will help us progress (conversely, you'd probably be wasting
> your time trying to illuminate the problem for me). So let us
> agree to differ, and yet be clear on the difference:
>
> o I do not accept that the CR argument proves *anything*: in
> particlar, it does not prove the existance of a "Strong
> Grounding Problem (SGP)" - i.e. that *no* system, can,
> soley by virtue of instantiating some particular formal

> program, achieve "genuine" understanding (or semantics), in the
> same sense that I (and, I conjecture, you) do.
>
> Equally, of course, CR does not *disprove* SGP either (that is
> to say that I am, for the moment at least, personally agnostic
> on this issue - though I promise to get off the fence and say
> where I'll actually lay my money, before the end of this message).
>
> o I *do* accept that there does not exist today, on this planet,
> any system for which it may be demonstrated that it, soley by
> virtue of instantiating some particular formal system, achieves
> genuine understanding. (Actually, I'm perfectly happy to grant
> the much stronger assesrtion that, as yet, the *only*
> things we know of which understand, in anything comparable to
> the sense way we do, are ourselves, and that it has
> not been demonstrated that we do this soley by instantiating
> some formal program - but I don't think this actually adds
> anything!)
>
> I call this (surprise, surprise) the "Weak Grounding Problem"
> (WGP), and I consider it a most serious problem confronting AI
> (but also connectionism!) - certainly insofar as these purport
> to offer theories of cognition, but also (like you I think) I
> am pessimistic about the prospects even for substantial
> progress in "applications" if the WGP cannot be effectively
> tackled.
>
>
> So? So, even though I disagree with you about the exact nature of the
> problem, I agree that there *is* a problem, and I think that
> (maybe) we can avoid getting bogged down in a discussion of
> Searle, while still addressing a real problem.

Sounds like all you're saying is that no one has written a program that can pass the TT (symbols-only version), and if no one ever can, or if it's impossible, that's bad news for AI. So what? That isn't a principled position, it's just a status report. Both Strong AI and Searle are concerned with what follows IF a program can pass the TT.

> -----------------------------------

>
> 2. What is a Symbol System?
>
> (The "real" question is: what are the causal powers, or
> functional capabilities, or, indeed, limitations, of symbol
> systems?)
>
> I welcome this attempt to clarify what is, and is not, a symbol
> system: for me, at least, it is long overdue. Needless to say,

> now that you have done the hard work of drafting a definition, I
> will do the lazy man’s work of quibbling with it.
>
> Firstly, I tend to agree with Kube that (2) & (3) may be
> unfortunate in that they appear to exclude certain things we
> conventionally think of as symbol systems - but that aspect
> doesn’t worry me unduly. In fact, whatever about using (2) & (3)
> in the definition of symbol systems *in general*, I think they
> are probably necessary elements for any putatively *intelligent*
> symbol system. The "reflexive hypothesis", as I call my version
> of (2) & (3), is perhaps a little stronger still - it says not
> only that the rules are themselves explicitly represented as
> tokens and strings of tokens, but that they are *accessible* to
> manipulation by other rules, or even themselves. Maybe you
> considered this was implicit in (2) & (3)?

Actually, I consider the 2nd order stuff (awareness of one’s symbols, etc.) to be a complete red herring. Typically once someone has gotten hopelessly lost in a hermeneutic hall of mirrors of his own creation (as a result of overinterpreting symbol systems) the stuff that’s interpretable as 2nd order awareness is touted as the system’s crowning glory. In reality, if you knew how to create a system that was aware of ANYTHING in the first place -- say, it could experience the pain of a toothache -- then the 2nd order stuff would be absolutely trivial by comparison, just a piece of fine-tuning now that we know that someone’s home in the first place.

> Anyway, I discuss this
> more fully in a conference paper I presented last year - its not
> a paper I’m particularly happy with, but I’ll certainly send you
> a copy if you’re interested.
>
> I *am* concerned, however, about (7) & (8): that it must be
> possible to *systematically* interpret *everything* - the entire
> system, and *all* its parts (composite or otherwise).
>
> >From my reading of your paper, this element of the definition
> derives fairly directly (though not exclusively?) from Fodor &
> Pylyshn, particularly their 1988 paper in Cognition. Well, its
> about a year since I read that paper, so I guess its time I went
> back and had another look. In the meantime, let me continue on
> the basis of my fading memory of it.
>
> I think you see (7) & (8) as a useful way of driving a wedge
> between "symbolism" and "connectionism" - a principled way of
> distinguishing the two ("symbolism" isn’t a great word here, but
> I think you know what I mean).
>
> The thrust of your whole paper then, is to say that a system
> which is, in these terms, purely symbolic *or* purely
> connectionist will be unable to instantiate "genuine" semantics;

> and that, therefore, we should conduct research into systems
> which are neither purely symbolic nor purely connectionist, but
> are, in fact, some kind of "hybrid" of the two - and you go on to
> fill in some details of a particular hybrid proposal. Implicitly,
> systems which are *purely* symbolic are a dead end from the point
> of view of cognitive theory.

A purely symbolic system is ungrounded; a purely connectionist system is not systematically interpretable and decomposable. But the hybrid system I propose is not symbolic/connectionistic, particularly; it's hybrid symbolic/nonsymbolic, with the main nonsymbolic functions being: sensory transduction, analog transformations of the sensory projections, and innate and learned (neural-net-mediated) detectors of invariance in the sensory projection.

> But there's a significant issue getting obscured here. F&P see
> connectionist systems as a subset of symbol systems in general
> (i.e. they are the class of "crippled" symbol systems which
> lack, possibly among other things, properties (7) & (8)), whereas
> you see them as *disjoint* from symbol systems - and the notion
> of a "hybrid" only makes sense under your interpretation (if
> connectionist nets are, from the point of view of relevant
> functional capacities, merely a specially restricted kind of
> symbol system, then the hybrid is not a hybrid at all - its just
> one big symbol system, parts of which are less "powerful" than
> other parts).

This sounds garbled. I have condensed a quasiformal set of criteria for what counts as a symbol system from the literature. The category does not seem to admit of degrees. Something either is or is not a symbol system. F & P, besides talking about symbol systems, talk about "representational" systems; these are systems that are interpreted as representing something. Nets can be thus interpreted, but because they lack decomposability and systematicity, they are not symbol systems.

On the other hand, nets can be simulated by symbols systems (most are). As such, whatever net simulations can do, symbols can do. Another possibility is that nets can be trained to become symbol systems, even programmable ones. I am not concerned with these issue becauses it does not much matter to me whether or not nets are symbolic. What matters is what it takes to ground symbols.

> The question is: can purely symbolic systems exhibit
> intelligence?
>
> Now, here I use "symbolic system", not in your sense, but in the
> F&P sense which *subsumes* connectionist nets as a special kind of
> symbol system.

Too bad, because as a consequence we are talking at cross purposes. I do not think F & P consider nets to be symbol systems, otherwise it is not clear what they're claiming nets lack or are unable to do that symbol systems have and can do.

> In particular, I accept systems which, in whole or
> in part, are merely symbolic "simulations" of connectionist
> networks. F&P do *not* accept that connectionist networks have
> any (relevant) causal powers, or functional capacities, that
> cannot be duplicated in some system which is "fully" symbolic -
> i.e. symbolic in *your* sense of the word. I emphasise
> *relevant* causal powers here - certainly, a symbolic simulation
> of a neural net is not "the same thing" as a neural net. Its
> not actually "the same thing" as anything except itself.

This is all a red herring. Symbol systems can simulate anything, according to the Church/Turing Thesis (which I don't particularly reject), including airplanes and furnaces, so there's no reason to suppose they can't simulate nets and even brains. The trouble is that simulated airplanes don't fly, simulated furnaces don't heat, simulated nets don't do whatever is essential to being a real parallel distributed net (if anything) and simulated brains don't think. All this because symbol crunching, be it ever so interpretable and universal, is just symbol crunching.

> By "intelligent" I mean intentional, conscious, having genuine
> understanding, incorporating intrinsic meanings, or semantics
> etc. - choose your own wording. (They're not all necessarily
> the same, but, from where we are today, they all look about
> equally intractable!)

You keep talking about intractability in practice, but Searle's Argument and the Symbol Grounding Problem are concerned with intractability in principle.

> Now, of course, you admit that this is an issue:
>
> ... It is not even clear yet that a "neural network"
> needs to be implemented as a net ... in order to do what
> it can do; if symbolic simulations of nets have the same
> functional capacity as real nets, then a connectionist
> model is just a special kind of symbolic model, and
> connectionism is just a special family of symbolic
> algorithms.
>
> But, almost in the same breath, we have:
>
> ... there has been some disagreement as to whether or
> not connectionism itself is symbolic. We will adopt the
> position here that it is not, BECAUSE CONNECTIONIST
> NETWORKS FAIL TO MEET SEVERAL OF THE CRITERIA FOR
> BEING SYMBOL SYSTEMS, AS FODOR & PYLYSHYN (1988) HAVE
> ARGUED RECENTLY.
>
> (EMPHASIS added)
>
> Now I think this is (unintentionally) misleading. I know (?) you
> actually take the view that symbolic systems do *not* have (at

> least) the same causal powers as connectionist networks (which
> is fair enough), but it looks as if you're trying to attribute
> this to F&P - i.e. to the lack of properties (7) and (8) -
> whereas it is precisely contrary to their view. Indeed, stated
> this way, the whole idea is absurd: it would be to suggest that,
> by taking away conditions (7) & (8), i.e. *weakening* a symbolic
> system, you would somehow confer upon it *extra* functional
> capabilities, rather than taking away from its existing
> abilities!

You misunderstand my grounding proposal if you think it's just imputing special powers to nets that symbol systems lack. I am suggesting that there are ways to put together a hybrid system, made out of transducers and analog transformers, one that uses nets as a component for doing feature learning, and that such a system could ground a symbol system that uses the names of sensory categories as its (grounded) atomic terms. My symbols/nets contrast is not as simplistic as the one you describe. And yes, F & P are making distinctions between two kinds of representational systems, systematic/decomposable and not, and the former are symbol systems and the latter (nets) are not.

> In fact, your reason for thinking that symbol systems do not have
> the same causal powers as connectionist networks is nothing to do
> with characteristics (7) & (8) of symbol systems and/or F&R's
> argument - it rests soley on Searle's CR argument. Which again
> is fair enough (even though I disagree), and you have said it
> earlier in the paper. But I think the subsequent discussion
> (emphasis even) of conditions (7) & (8), as a demarcation
> between connectionist and symbolic systems, gives entirely the
> wrong impression; it implies that, hidden somewhere behind (7)
> & (8), lies the answer to Searle and to the (strong) symbol
> grounding problem itself. This is illusory, and I can't believe
> you *mean* to suggest this.

Nothing of the sort; in fact, early in this discussion group I aired a generalization of Searle's Chinese Room argument that applies as much to nets as to Strong AI. (It's a 3-room argument; in one room a real parallel-distributed net passes the TT, in a second a symbolic simulation of that same net passes the TT, and in the 3rd room Searle simulates the second; conclusion left as an exercise.) You have unfortunately interpreted my paper as merely discussing symbols vs nets, whereas the hybrid system I actually propose depends essentially on analog structures and processes, with nets merely doing feature learning.

> Let me state what I think is your position, and you can
> contradict me as necessary:
>
> o You are persuaded by Searle's CR argument that purely
> symbolic systems cannot have true understanding.

Correct.

> o You don’t know (any more than Searle does) exactly what the
> special causal powers are that people (or, more specifically,
> brains) have that mean they can support true understanding.
> You merely assert that, whatever they are, they are absent
> from purely symbolic systems.

And I hypothesize that they reside in a robot’s capacity to pick out the objects and states of affairs to which its symbols refer, on the basis of their sensory projections. And hence that the symbols are grounded in sensory projections, analog transforms of them, and invariances within them, perhaps learned by nets.

> o You conjecture that artificial neural networks (connectionist
> networks) *may* have some or all of these special causal
> powers, alone or in conjunction with other kinds of system
> (such as symbolic systems) and that systems incorporating such
> networks may, therefore, be capable of true understanding.

You leave out all the transduction and analog transformations I believe are essential in all this. I just use nets to learn the invariant features of sensory categories.

> o BUT (you admit): this can only be so, if those neural networks
> have some relevant causal powers that *cannot* be duplicated by a
> symbolic system.

Not at all; I don’t care whether the nets are real or simulated, as long as they can learn the invariances that will reliably categorize sensory projections and allow their distal objects and states of affairs to be named. I don’t talk much about "causal powers" at all -- except in reminding hermeneuts that simulated planes can’t fly and simulated brains can’t think.

> o Now, as far as I can see, all the things that you propose
> for implementation by connectionist network(s) in your
> hybrid model, could be functionally duplicated by a suitable
> symbolic system simulating the network(s) - i.e. literally, the
> neural network components could be unplugged, replaced by
> suitably programmed computers, and the system would continue
> to behave as it did before. There would, of course, be
> problems in building computers that were fast enough and/or
> large enough, not to mention writing the software,
> but there is no problem *in principle* (that I can see).

True, but completely uninteresting, because nothing hinges on it. A potentially more consequential objection has been attempted by others, in which it is suggested that all the transduction and analog processes inside the robot, as well as its world, could all be symbolically simulated too. Unfortunately, all that gets you is a simulated robot in a simulated world, which should no more be expected to think than a simulated plane whould be expected to fly. The only system that can think is a system that that has REAL grounding, not simulated grounding. It must pass the real TTT, for the "simulated" TTT is just the TT and Searle all over again. Simulated grounding is hanging by the same skyhook as the original symbol grounding problem itself!

> o Therefore, in proposing your hybrid model, as a resolution of
> the (strong) Symbol Grounding problem, I think there is a
> clear onus on you to point to the specific function or
> functions which you are postulating in your system that could
> *not* be implemented (NB: I mean *implemented*) by a suitable
> symbolic (sub-)system (which, we may say, is *simulating* some
> other implementation - but where any such distinction is functionally
> opaque to the rest of the system), and explain just *why*,
> even in principle, this function could not be achieved by a
> symbolic system.

Transduction of the proximal projection of the distal object, and all internal analog transformations thereof, including reductions to invariances within the sensory projection.

> o Now I don't think you can do this (though I'd be very
> interested in any attempt you might make); but then, that's
> clearly because I don't see the merit in the CR argument in the
> first place, and therefore don't admit, yet, that an SGP
> exists. Indeed, the challenge is just a variation on one of
> the challenges commonly put to Searle. The difference in your
> case is that you cannot hide behind our ignorance of real
> brains (which is certainly profound): your system is of your
> own designing - there's nothing there except what *you* put in.
> If you havn't put in the magical biogunk, it ain't there.

Quite correct, and as you see, I'm not hiding. By the way, Searle's Chinese Room is just an instance of the Chinese Room Problem; it is not the sole or even the principle argument for it. For example, reread the section on the Chinese/Chinese Dictionary-Go-Round in my paper.

> o Notwithstanding my pessimism on this point, I *don't*
> propose to disregard or ignore your proposed architecture,
> because of that. What I mean is that, even though your system
> may not solve the SGP (perhaps because there is no such problem
> and therefore it cannot be solved), it may still be a tentative
> solution to the WGP - and we can proceed to consider it on its
> merits, in *that* light. Well, at any rate, that's how *I*
> intend to proceed.

There is no "Weak Grounding Problem." As I said, that's just a status report on what AI modelers have not yet managed to do so far, not a principled analysis of why or why not.

> -----------------------------

>
> 3. Your Hybrid Model.
>
> This is the real meat of my comments. It may appear rather radical.
>
> I follow Karl Popper: I don't believe in induction.

>
> I have no question mark in my mind about the "inductive power of
> neural nets" - they don’t have any (nor, for that matter, do
> existing so-called Artificial Intelligences).
>
> OK let me be more explicit.
>
> We agree (I hope) that there is no such thing as "inductive
> logic" - i.e. that no set of particular facts (no matter how large)
> ever logically entails a generalisation of those particular
> facts (no matter how often the Sun rises in the morning, it does
> not logically entail that it will ever do so again etc.).
>
> Indeed, for any given set of particular facts, there are always
> an infinite number of different generalisation which are
> compatible with that set of facts.
>
> But, you say, a set of particular facts may *imply* (more or
> less "strongly") some particular generalisation(s) over others?
>
> Nope, afraid not: the particular facts, in themselves, cannot
> yield a probability distribution over the (infinite) set of
> compatible generalisations. They do not make any one (of the
> consistent) explanations more "probable" than any other.

What is all this about? I don’t talk about the philosophical problem of induction, just the real problem of sensory category learning, a problem there is every reason to believe IS soluble, since real organisms seem to solve it quite successfully. Nor is there a "weak induction problem" corresponding to the fact that we do not yet have a successful model for learning. It is of course logically possible that we will never find such a learning model (just as we may never write a program that can pass the TT), and even that organisms can’t learn either, and that all knowledge was already inherent in the structure of the Big Bang. You’ll forgive me if I don’t give that possibility much credence until it looks absolutely unavoidable...

> What particular facts can, and do, do is to *rule out* (refute)
> certain generalisations - i.e. all those generalisations which
> are incompatible with one or more of the particular facts. And,
> as more particular facts are encountered, more and more
> generalisations may (in principle at least) be eliminated - but
> remembering all the while, of course, that we shall never, in
> this way, find the "true" or "correct" generalisation, for there
> will always remain an infinite number of candidates (infinity
> less any finite number is still infinity, unfortunately).

In another paper I elaborate on the approximatateness of category learning, but so what? Why all the apocalyptic talk about "truth"? We’re just trying to model what organisms can and do do, not what they can’t (because nothing can).

> Does this mean that there cannot be *any* basis for prefering
> one generalisation to another?
>
> No, not at all - indeed, we (and systems in general) do it all
> the time. We guess, we conjecture, we hypothesise, we invent,
> we suppose, we imagine etc. etc. And (if we are remotely
> adaptive) as new particular facts come in, we modify our
> positions (our theories, if you will). This process (which you
> may call induction, if you wish, but it bears very little
> resemblence to the concept of "logical" induction) is, in
> different systems, more or less "inspired" or effective - that
> is, different systems seem to be better than others at
> formulating improved theories.
>
> Man, in particular, excels at it, and has discovered what appears
> to be one particularly good methodology (heuristic, if you will):
> having formulated a theory, try your best to refute it (both
> through finding "internal" or logical problems with it - i.e.
> discovering that it isn't really consistent with the particulars
> you already know of - and through seeking out new particular facts
> that may decide between alternative theories). This is science.

This is all quite vague, but to the extent I can put a gloss on it, it sounds like ungrounded bottom-uppism all over again: For wherein is the MEANING of the "theory" to reside if its terms are but ungrounded squiggles and squoggles?

> But be careful: I have not (and do not pretend to have) defined
> any (specific) effective procedure, or computable function here.
> Firstly I havn't addressed the issue of formulating a new theory
> in the first place. But even the question of seeking refutations,
> formulating crucial tests etc., is not generally some unique
> "logical" process.
>
> That is: you can adopt procedures for selecting "preferred"
> generalisations from particular facts, and some will be "better"
> than others. But, since there is no "correct" procedure, any
> such procedure should be tentative. In effect, at any
> particular level in this hierarchy, the procedure you use for
> generalisation ("induction") is, itself, an hypothesis, a theory
> (about how to do induction) which is fallible and (in the more
> adaptable system) should be continuously subject to attempted
> refutation - or whatever other higher level procedure may be
> formulated.
>
> This is very close to saying that "intelligence" cannot be
> captured in any effective procedure (program) - though on
> grounds entirely distinct from Searle. It is closer to the
> argument put forward by Lucas, in the context of Goedel's

> theorem, though it is still quite distinct.

It also sounds to me extremely vague and figurative. It’s not clear what’s at issue, and what follows from what, and why. Or what its relevance is to the real problem of learning and categorization in organisms.

> Lest there should be any confusion therefore, let me make it
> clear that I am *not* claiming that intelligence cannot be
> captured in an effective procedure. Indeed, I hypothesise that
> there exist effective procedures which, if suitably instantiated,
> could manifest intelligence of entirely the same sort as we do
> (the instantiation *is* important - a disembodied, or quiescently
> archived procedure would certainly have a more dubious status).
> That is, I subscribe (on a tentative basis) to the so-called
> Church-Turing thesis, or even the physical symbol system
> hypothesis, as near as dammit. (I promised I’d come down off the
> fence, even if only tentatively...)

Sounds like you subscribe to several of the premises that have by now been shown to lead to untoward conclusions. Unless you merely wish to ignore those untoward conclusions, you will have to change your position to accommodate them. You should find, if you think about it, that ungrounded symbol systems can no longer be rationally expected to accomplish and instantiate what you had thought, despite the Church/Turing Thesis.

> But, I think that such systems could be ranked reasonably
> objectively on a scale of intelligence - you know, rocks at the
> bottom, thermostats still not really registering, the most
> advanced, adaptive, dedicated artificial control systems might
> just be discernable above this level, then a big gap before we
> come to the simplest biological systems - viruses say - on up
> through bacteria, the simplest multicellular organisms, plants,
> insects, vertebrates etc., with man at the current peak (?). My
> biology is lousy, but I hope you get the idea of a ranking based
> on adaptability - on the ability to formulate, effectively and
> efficiently, "good" theories about the world (i.e. better than
> the ones generated by the systems lower down on the scale). The
> point is that, because there is no such thing as induction - no
> such thing as an effective procedure for induction - there can be
> no absolute maximum on this ranking. There may (in theory at
> least) always be a better system, as ranked on this scale.
> Conversely, there may be no "most intelligent" system, or, even
> if there were, we wouldn’t know how to recognise it (God does not
> exist? Ooops, I’d be getting into very deep water there....).

The question of whether consciousness can be seen on a continuum has already been much debated in this group and elsewhere. The view from here is that the idea is incoherent. So there’s no 1st order continuum. There’s either somebody home or not. As to a quantitative 2nd order continuum: It’s of no interest whatsoever, because either you have the 1st order stuff (for some other reason, yet to be explicated), in which case the higher-order stuff is just the frills and fine-tuning, or else you don't have the 1st order stuff, in which case all the stuff that can be interpreted as intelligence, etc., is just hanging from a skyhook. Or, in the inverted metaphor: It's utterly ungrounded.

> You'll notice I didn't even mention AI programs or connectionist
> networks on that ranking. That's because, of the ones I'm
> familiar with, none seem to have made any serious progress at
> realising the kind of adaptability I'm talking about here. That
> may be unduly harsh: take it simply as a value judgment that such
> progress as there has been made is small in the context of the
> problem. This is no criticism or insult to the exceptionally
> talented people who have worked in the field - merely a personal
> perspective on the scale of what we are attempting to do!

What you have said has also been extremely vague. AI and connectionism are at least explicit about where they pin their hopes: One on symbol crunching and the other on, say, hidden units and the delta rule. You've just spoken of "theories" and increasingly "efficient" formulators of them. Where does that get us?

> So what's the point of this monologue?
>
> Merely to challenge you to respond. I claim that the kind of
> system you have sketched out (your hybrid connectionist/symbolic
> system) though directed at the right problem (WGP - which I take
> to be a version of the "heuristic" induction problem) doesn't
> have the right structure about it - it "feels" wrong. Its
> probably an improvement over many of the exclusively connectist,
> or exclusively symbolic, systems that are being built, but I
> still don't see that it has really scratched the surface of the
> problem of how certain very special organisms (us) can develop,
> in a relatively short time, extraordinarily complex and effective
> techniques for modeling the reality we are confronted by (which
> is just another way of saying we construct, among other things,
> symbol systems with "deep" meaning, effectively, though not
> definitively, grounded in our shared reality).

I agree that it's just scratched the surface and will probably, like all early hypotheses, be wrong. (I don't yet see why it "feels" wrong.) So what? Alternative solutions to the symbol grounding problem will have to be at least as specific; and will have to try to SOLVE it, not just deny or ignore it.

> In particular, the only "grounding" processes you come up with
> seem to be based on experiments in training networks to
> recognise, or discriminate patterns. As far as I am aware
> (I could certainly be wrong here) such recognisers are generally
> the very antithesis of the kind of adaptive system I'm
> hypothesising. They work, not by bold conjectures and seeking
> out refutation, but, faced with refutation, making the minimum
> adjustment that is consistent (if even that much). More
> seriously, they lack any metatheory - any basis for criticising

> their own so-called induction, and trying to improve it. They
> lack the many, many layered, hierarchical structure (I'm talking
> functional, not physical, structure) that would characterise any
> highly adaptable system. Critically, and this is my real
> complaint about work to date under the banners of AI and
> connectionism, and even (to to more muted extent) cybernetics:
> they lack any apparent mechanism whereby the hierarchical
> structure could *grow* - could elaborate itself, really learn
> from its experiences, and especially failures, in the world.

> Thanks, Barry.

Mine is a bottom-up grounding scheme; if it works, your theories and metatheories will come later; first let's get a robot that can make its way in the world and pick out the categories of sensory experience that matter; once it can name those, it can systematically string names into composite decsriptions of more abstract objects and states of affairs. That may be the level at which your theories (grounded in and hence constrained by this substrate) come in.

Stevan Harnad

TUTORIAL ON THE REAL/AS-IF DISTINCTION

> From: Jeff Inman
>
> I am certainly convinced that "AI"-type symbol crunching has little to
> do with "meaning", in the sense you mean. However, I think you are
> trying to make an axiom out of Searle's arguments about the Chinese
> Room, claiming that the Chinese Room was proved to have no "understanding".
> I for one am not yet convinced. I suppose that angers you because you
> feel (perhaps rightly) that I MUST simply be too dense to absorb your
> arguments against the "systems response". Much as I am impressed by
> the forcefulness and sophistication of your arguments, I haven't yet
> seen a neccessary argument against the systems response.

There is a simple reason why you will not see such a "necessary" argument: Because of the (insoluble) other-mind problem it is impossible to "prove," as a matter of necessity, that anyone or anything does or does not have a mind. It's not a question that can be settled with the force of necessity, and the only one who knows for sure whether there's anybody home in any system is the one who is (or is not, as the case may be) home. It's as impossible to "prove" that a system has a mind (or doesn't) as it is to prove that there is a God, or isn't (or that there was Special Creation, or what have you).

For that matter, even F = MA does not have the force of necessity, but there the preponderance of the evidence is good enough. What is the evidence in the case of Searle's Argument? There was this hypothesis, that thinking is just symbol crunching, and that a system thinks in virtue of implementing the right symbol system. It looked good on the face of it, for various reasons (the successes of AI, the power of the Turing Machine, the apparent fit between the software/hardware distinction and the mental/physical distinction), and then Searle pointed out that, if this hypothesis were correct, it would follow that he too should understand Chinese if he implemented the right symbol system. Since he could do this by merely memorizing how to manipulate a bunch of meaningless symbols, it was clear that the hypothesis led to absurd conclusions and could therefore be rejected.

What do those who are "unconvinced" reply: That memorizing a bunch of symbols would cause Searle to develop another mind, one of which he was unaware, but one that did indeed understand Chinese. As I've said countless times, this "Systems Reply" is merely a reiteration of the hypothesis itself ("to think is just to implement the right symbol system"), louder, and at a higher counterfactual price. (This response is of course always possible, because the hypothesis can never be shown to be false as a matter of necessity, as I noted.) In this case, we are to believe that memorizing symbols has a new, unexpected power, despite the utter absence of evidence or any other reason to believe it, except the hypothesis itself, which we are trying to rescue from its absurd consequences. If this isn't theory-saving I don't know what is. But I'm sure one could always rescue Special Creation by exactly the same kind of manoeuvre....

> The case, as I
> see it, hinges on what we mean by "the system". Your approach seems to
> take the standard view that science consists of understanding
> loosely/un-coupled subsystems which exist together, interacting in a
> vacuum. Thus, you posit people as self-contained, autonomous agents.
> This shows up in your stance on the other-minds problem, and in your
> development of the TTT(T).
>
> The advantage to your position is that you have enough information to
> make your arguements. As a systems responder, I lack enough information
> to define the meaning of a symbol because I consider that I myself am a
> component in the system, and so my "understanding" is really part of
> the physics of the system. The meaning of anything is SYSTEMIC. At this
> level, a question about the mindfulness or mindlessness of a localized
> being is not a test for the presence of "meaning".

But you see, I believe in systems too. It's just that symbol systems are not the only systems I believe in, and in particular, I think (for many reasons, including Searle's) that it is by now clear that whatever kind of system a system with a mind is, it is not just an implemented symbol system. Systems-repliers don't have a monopoly on systems. My proposed hybrid nonsymbolic/system system is a system too, but it has the virtue of being immune to Searle's Argument and being grounded in the world (in virtue of the TTT), as we are.

> What you must mean is that there is an essential difference in the
> energy being transduced. No argument there. Agreed; they are different
> processes. One of them you attribute meaning to, the other not. Do not
> mistake me as saying they represent [or mean] anything at all similar,
> because I'm not. However, if we look at these two phenomena with our
> eyes sort of unfocused, trying get their GIST as physical phenomena,
> then I don't see any difference between them. The attribution of "as
> if" requires a context .. you need to say "as if WHAT, in WHAT sense,
> to WHO, WHEN"? If then you say "like what I thought they meant when I

> saw them", you have narrowed the context to make a distinction, but you
> have not proven that they are somehow higher or lower class phenomena,
> except by some unnecessary criterion.

The real/as-if distinction is perfectly clear and unproblematic. First practice on the distinction between a real plane, that really flies, versus a symbol system that is merely interpretable as if it flies. Then move on to furnaces, real and symbolic. Once you have the distinction well in hand, move on to transducers, real and simulated, and finally to the distinction between systems capable of passing the real TTT vs. simulations of them (which just amount to the symbols-only TT all over again). Systems capable of passing the real TTT cannot be just symbol systems, and you don't even need Searle's argument to see that.

> However, I'd argue that REAL meaning occurs at the System level.

Fine, but what kind of System? The point here has been that there's no meaning in an implemented symbol system that is interpretable as meaning, just as there is no heat in an implemented symbol system that is interpretable as heating.

> To me, the distinction above is really a definition of two types of
> mindfulness. You want to distinguish someone who really knows what they
> mean from someone who is just pretending that they do (like me). I'd
> have to ask you [as I'd ask anybody]: how can you tell that what you
> think you mean is really what you mean? Your answer is that, whatever
> you think you mean is what you call meaning! Okay. Now, carefully,
> let's look at the mindless robot. Let us suppose that it is faithfully
> transducing signals just as you were, but lacks what you recognize as
> an experiential state that you would relate to the one you had, when
> you were saying similar things. (1) Can you be sure that these two
> mindfulnesses are necessarily completely unrelated, whereas any two
> conversing people must have more "in common"? And (2) even if you
> could, (and I might agree with you), is there no spectrum on which both
> types of mindfulness can exist?

First of all, you have missed a critical point: According to me a TTT-capable robot WOULD have a mind, meanings, etc., but it wouldn't be just a symbol cruncher either. So you have a straw man here. Second, as I said, even the TTT cannot have the force of necessity, because of the other-minds problem. Finally, until further notice, it is no more true that there are two kinds of mindfulness -- real, and symbolic-but-interpretable-as-if-mindful -- than there are two corresponding kinds of flying.

> Ultimately, I argue that we can carry this arguement all the way to a
> console that is displaying symbols. However, I also claim that the
> console (or mindless/mindful robot) itself is not the complete system,
> nor is the meaning that you perceive the complete meaning. You may have
> anticipated what I mean by "the system" by now; it is nothing less than
> everything. In fact, this is also the only REAL system that is
> complete. Otherwise, if you want systematically interpretable, you'll
> have to beware of who's doing the interpreting.
>

> They are "as if" in terms of the meanings you'd be inclined to give to
> them. In terms of the meanings they actually have, they can hardly be
> anything except REAL. Jeff Inman

This has lapsed into the kind of holistic mysticism I cannot follow ("we're all parts of cosmic symbol system in the mind of God, or what have you"). I'll have to get off the starship at this point...

Stevan Harnad

TURING EQUIVALENCE IS NOT GROUNDING

> From: kp@uts.amdahl.com (Ken Presting)
>
> KP: Glad to see the group back on the air! Thanks for your remarks on my
> suggestion. I'll elaborate and focus a little more:
>
> > >(Ken Presting:)
> > > Note first that the symbolically encoded rules given to Searle are
> > > identical to the program executed by the computer (if they aren't,
> > > they might as well be). Now, Searle reads the rules, understands
> > > them, and intentionally obeys them. Therefore the rules are *not*
> > > devoid of semantic content, and Searle's Axiom 1 (Sci.Am. version)
> > > is false.
> >
> >(Stevan Harnad:)
> > No. The rule says, if you see a squiggle, give back a squoggle.
> > The foregoing sentence has semantic content all right, and Searle would
> > understand it, but it's not THAT semantic content that's at issue.
>
> KP: You are certainly correct that there can be no issue of *whether* the
> rules have any semantic content. But I think there may be an issue
> regarding the nature of the semantic content of the rules. >
> > SH: In the Chinese case, it's the meaning of the Chinese symbols (because
> > that's what Strong AI-ers are claiming that the system which is
> > implementing the program is really understanding) . . .
> > THAT's the semantics which is at issue; not the semantics of the
> > squiggle-table (which, by the way, Searle both understands AND
> > implements, whereas the computer only implements).
>
> KP: It may help to propose a distinction between the meaning of the Chinese
> symbols as they appear in the rules, and the meaning of the same
> symbols as they appear in the inscriptions of a Chinese author. In the
> case of the rules, the Chinese symbols denote their own shapes, whereas
> in the case of a Chinese inscription, the symbols have the usual sense
> and denotations of horse, stripes, etc. (To avoid difficulties with the
> theory of quotations, we could replace the occurences of Chinese
> symbols in the rules with English or PostScript descriptions of the

> shapes).

I think you missed the point. The rule: "if you receive squiggle, send squoggle," does have semantic content, and Searle does understand it, but it's not Chinese, and it's the understanding of Chinese that's being tested in the TT. It won't do to say "well, at least there's SOME semantics going on then" because no one claimed that the computer's "understanding" consisted in "Ah, when I get this symbol, I'm supposed to send that one." It's the semantic description of the uninterpreted syntactic rules that Searle of course understands, because you have to explain SOMETHING to him to get him to do anything, given that he is, after all, a conscious person. But that does not imply that the COMPUTER understands the syntactic rules any more than it understands their semantic content. In any case, besides its obvious absence in the computer, the point is that it's the wrong semantics.

I also hope you are not making the mistake of thinking that because the syntax is semantically interpretable the semantics is therefore intrinsic to it: A book is already a counterexample to that. Its symbols are ever so interpretable as meaning what we mean by those symbols, but the book itself does not "mean" anything. By the same token, neither does the computer, despite all the interpretable syntax. Least of all does it mean: "I am manipulating squiggles and squoggles thus and so..."

Your suggestion that "In the case of the rules, the Chinese symbols denote their own shapes" is misusing the word "denote." In a machine implementing the statement "If read '0' print '1'," the symbols don't "denote" their own shapes. The description of what the computer is doing merely denotes what the computer is doing -- for us, not the computer. The computer is just doing it. Forget about computers. In the uninterpreted syntax of Paeano arithmetic, "0' = 1" does not "denote" "0' = 1." It's just a string of symbols interpretable as meaning (denoting) "the successor of zero is one."

> KP: It is crucial to the next step of my argument that the computer does
> only implement the rules, so I'm glad that you have noted that point.
>
> > > KP: Grounding comes in when the rules are loaded into a real computer,
> > > and the question of the semantic content of the rules arises again.
> > > Now the rules, as they exist inside the machine, determine the
> > > causal process of the machine's operation. The correspondence
> > > between the syntax of the rules and the operation of the machine is
> > > precise (for obvious reasons). Therefore the symbols used in the
> > > rules *are* grounded, and this holds for every program.
> >
> > SH: I'm afraid you've misunderstood the grounding problem. To implement a
> > program (which is all you're talking about here) is decidedly NOT to
> > ground its symbols. If it were, there would be no grounding problem at
> > all! No, the symbols are also presumably semntically interpretable as
> > meaning something more than "If squiggle then squoggle."
>
> KP: Let me check my understanding of the SGP by attempting a rough
> statement of it: "Under what circumstances can a unique interpretation
> be assigned to the linguistic performances of a system?" The central
> application for this question for AI is of course to the inscriptions

> output from a computer.

That's not the Symbol Grounding Problem! That a unique interpretation can be assigned to syntactic performance must be true if a system is to qualify as a symbol system in the first place! But it is symbol systems that have the symbol grounding problem, which is "How can the interpretation of the symbols be made intrinsic to the system itself, as the interpretations of the symbols in our minds are, rather than parasitic on the interpretations projected by our minds on the symbol system." There is no question about the EXISTENCE of unique interpretability; it's its GROUNDEDNESS that's at issue. My own candidate solution happens to be that symbols are grounded in the objects and states of affairs they refer to by the causal wherewithal to pick them out from their sensory projections (the Total Turing Test). This depends critically on nonsymbolic structures and processes like transduction, analog transformations and sensory invariance extraction. A grounded system must therefore be hybrid nonsymbolic/symbolic. A pure symbol system is ungrounded and therefore cannot be a model for the mind, cannot really understand (be it ever so uniquely interpretable as understanding), etc.

> KP: However, the question also has an interesting application to a certain
> subset of the inscriptions produced by programmers - our programs! When
> Searle's programmer's write "If squiggle then squoggle", what does that
> mean? It means that when a squiggle comes in, a squoggle had better go
> out, or else it's time to call the repairman! That is, an
> implementation of the program must instantiate "If squiggle then
> squoggle" as a lawlike consequence of its causal powers.
>
> This defines an equivalence class of physical devices, as long as the
> typography of squiggles and squoggles is held constant. Now, it is very
> important for the theory of algorithms and for the philosophy of mind
> that the typography of public languages is as irrelevant to computation
> as it is to thought. But, noting that both fields have a similar
> indifference, I think we can standardize on Pica (or whatever standard
> exists in Chinese) without loss of generality.
>
> The upshot is twofold: every symbol in the syntactic rules is grounded
> (with the Chinese symbols grounded trivially in their shapes), and an
> implementation of a program does have specific causal powers. At least,
> every machine in the equivalence class of implementations of a given
> program shares the causal powers which entail its adherence to the I/O
> conditions of the program.

You've gotten out of your deduction exactly what you put into it: You have an equivalence class of algorithms that have the same semantic interpretation. Fine. But this does not help the philosophy of mind unless that interpretation is grounded. To put it another way, the software/hardware distinction, even though it looked promising for a while, does NOT capture the mental/physical distinction. It's harder than that.

> > > KP: A few disclaimers. I'm not arguing (here) that either the CR or
> > > the computer understand Chinese, I'm only arguing that Searle's
> > > argument fails. Also, I'm making no claims about the groundedness
> > > of any Chinese symbols as used by the CR or the machine, the only

> > > groundedness claim is for the symbols used to state the rules. I'd
> > > like to take up these issues later, however.
> >
> > SH: But no one was particularly concerned with the statement of the
> > syntactic rules. If Searle didn't understand those then the only way he
> > could implement the program would be to have someone or something else
> > PUSH him through the motions! From Strong AI's standpoint, it's enough
> > that he is actually going through the symbol manipulation sequence,
> > i.e., he is indeed an implementation of the program that putatively
> > understands Chinese. If you're not interested in whether or not he
> > understands Chinese then you've changed the subject.
>
> KP: What I'm suggesting is that by attending to the groundedness of the
> syntactic rules, we can learn something about the causal powers of
> programmed computers. I see two issues wrt causal powers. One is the
> general issue of whether running a program on a computer brings it about
> that the computer has any specific causal powers. Second is whether
> there is any program such that by running it, a computer can have
> causal powers equivalent to the brain.

But groundedness does not mean "uniqueness off semantic interpretability," or "equivalence class of algorithms." It means the symbols really mean what they mean IN THE (putative) MIND OF THE SYSTEM ITSELF, not merely because they are interpretable as meaning what they are interpretable as meaning, or because our minds interpret them as meaning what they are interpretable as menaing, but because they really mean what they mean. To put it another way: There's really somebody home in there, just as there is in here. Nobody home, no thinking.

A better distinction to keep in mind is the distinction between the real flying a real airplane and the simulated flying of a symbol system that can be interpreted as flying. Turing equivalence is irrelevant. One flies and the other doesn't. The same is true of real systems that really think versus symbol systems that can be interpreted as thinking. We, for one, are systems that really think; I'm proposing that the right "equivalence class" is not Turing Equivalence but the causal capacity to pass the TTT. That is my proposed litmus test for whether or not the thinking is real. The TTT is still fallible, of course, because of the (insoluble) other-minds problem (the flip side of the mind/body problem, likewise insoluble), but at least a TTT-scale system can break out of the symbolic circle to which pure symbol systems are doomed, in that a TTT-scale system's symbols are grounded in the causal power to pick out their real referents. That's the only relevant causal power here, not the causal power to generate syntactic performance that is interpretable as if... etc.

> At this point I would be satisfied to get agreement that when a
> particular program is run on a particular computer, the computer does
> have some specific causal powers. What those powers are, how they are
> related to the program, and whether the powers can be equivalent to the
> brain's, must be discussed afterwards.
>
> Ken Presting

Of course it has "causal power" -- whatever that awkward phrase means, every physical system has that -- but unfortunately it does not have the right causal power for implementing a mind, any more than it has the right causal power for implementing a plane.

Stevan Harnad

Synopsis of the Symbol Grounding Problem

Stevan Harnad Department of Psychology Princeton University Princeton NJ 08544

A symbol system is a set of physical tokens (e.g., scratches on paper, holes on a tape, flip-flop states in a computer) and rules for manipulating them (e.g., erase "0" and write "1"). The rules are purely syntactic: They operate only on the (arbitrary) shapes of the symbols, not their meanings. The symbols and symbol combinations can be given a systematic semantic interpretation, for example, they can be interpreted as meaning objects ("cat," "mat") or states of affairs ("the cat is on the mat"). The symbol grounding problem is that the meanings of the symbols are not grounded in the symbol system itself. They derive from the mind of the interpreter. Hence, on pain of infinite regress, the mind cannot itself be just a symbol system, syntactically manipulating symbols purely on the basis of their shapes. The problem is analogous to attempting to derive meaning from a Chinese/Chinese dictionary if one does not first know Chinese. One just goes around in endless circles, from meaningless symbol to meaningless symbol. The fact that the dictionary is systematically interpretable is of no help, because its interpretation is not grounded in the dictionary itself. Hence "Strong AI," the hypothesis that cognition is symbol manipulation, is incorrect, as Searle has argued.

How can one ground the meanings of symbols within the symbol system itself? This is impossible in a pure symbol system, but in a hybrid system, one based bottom-up on nonsymbolic functions such as transduction, analog transformations and sensory invariance extraction, the meanings of elementary symbols can be grounded in the system’s capacity to discriminate, categorize and name the objects and states of affairs that its symbols refer to, based on the projections of those objects and states of affairs on its sensory surfaces. The grounded elementary symbols -- the names of the ground-level sensory object categories -- can then be rulefully combined and recombined into higher-order symbols and symbol strings. But unlike in a pure symbol system, these symbol manipulations would not be purely syntactic ones, constrained only by the arbitrary shapes of the symbol tokens; they would also be constrained by (indeed, grounded in) the nonarbitrary shapes of the distal objects, their proximal sensory projections, the analogues of the sensory projections that subserve discrimination, and the learned and innate sensory invariants that subserve categorization and naming.

The solution to the symbol grounding problem, then, will come from an understanding of the mechanisms capable of accomplishing sensory categorization and the learning of concrete and abstract categories. Among the canidates are sensory icons and neural nets that learn sensory invariants.

> Date: Sat, 4 Aug 90 15:10:53 EDT
> From: block@psyche.mit.edu (Ned Block)
>
> I'm unclear on how you differ from searle on the relation between conscioun
> ess and intrinsic intentionality. He says POTENTIAL conssciousness is
> necessary for ii, and you say ACTUAL consciousness is necessary for ii?
> That couldn't be it--could it?
> Ned

Here is precisely the way I would put it: The only difference between a system that has intrinsic intentionality (i.e.. really means, thinks, etc. that X) and a system that has only derived intentionality (i.e., is systematically interpretable as meaning, thinking, etc. that X, but in reality does not mean, think, etc. at all) is the fact that there is something it is like to mean, think, etc. that X. I.e., there's got to be somebody home.

I don't care much about the potential/actual consciousness distinction insofar as the specific state of meaning X is concerned. Perhaps most of its neural substrate is conscious, perhaps it's not. Perhaps the subject is aware of meaning X, perhaps not. It doesn't matter. My point is simply that there is no tenable real vs. as-if distinction except with respect to whether there's really somebody home.

Here is an even more intuitive way to put it: I challenge you to formulate a coherent real/as-if (or intrinsic/derived) distinction if it is GIVEN a priori that the two hypothetical (turing indistinguishable) candidate systems in question (one with real aboutness, one without) are BOTH not conscious, never were, never will be; i.e., with the premise that there's nobody home in either case. Apart from voodoo, there's just nothing left of the difference between real and as-if aboutness on that assumption!

I think the foregoing is just obvious. But then why isn't it just as obvious that it's consciousness that's doing the work in the intrinsic/derived distinction all along? In other words, that as a "mark of the mental," over and above consciousness itself, "(intrinsic) intentionality" is utterly redundant. In my view the bifurcation between conscious states and intentional states, and the hope that the latter could be independently treated on their own terms, was utterly wrong-headed, despite all the serious philosophical store that has been set by it since Brentano et al.

That's the point on which John and I continue to disagree .

Cheers, Stevan

> Date: Sun, 5 Aug 90 22:16:30 EDT
> From: block@psyche.mit.edu (Ned Block)
>
> Stevan, in not caring about the potential/actual consciousness
> distinction, you are not distinguishing between 2 very different
> claims. I suspect that you may not really believe one of them and if
> you do, it certainly is not the important difference between you and
> John. The other is a claim that I don't think john would disagree with

> at all, so your note to me doesn't get at whatever you and John are
> disagreeing about.
>
> Your first claim, the one I suspect you don't want to assert
> confidently on reflection is that for me to really think or desire that
> X, there has got to be something it is like for me to think or desire
> that X--X has to have phenomenal consciousness. This is what is stated
> in the first paragraph of your message to me. The problem is: what if
> my desire that X is UNCONSCIOUS? Then you would have to hold that
> either (1) there is always something it is like to have unconscious
> desires (something I am willing to contemplate, but I doubt is one of
> your commitments) or (2) unconscious desires aren't cases of intrinsic
> intentionality.

I suspect that it's true that to really think or desire X there has to be something it's like to think or desire X, and that the same is true of unconsciously thinking or desiring that X, i.e., that whereas the latter is not like the former, it is nevertheless like SOMETHING (as opposed to nothing).

But these are all just very vague quibbles about the fine-tuning of particular states -- particular configurations of the internal parts and goings-on of some unspecified actual system or class of systems. What you are isolating below as the second claim is, in my view, really the heart and soul of the matter. For the SUBSTANCE of what the physical realization of mental states consists in will surely depend on what kind of system a system has to be in order to be able to be conscious at all. Once that's settled, the rest of the mind/body problem will be a piece of cake. In particular, whether all intrinsically intentional states configure the components (which have already been picked out as necessary for consciousness by the answer to the substantive question) this way or that way (e.g., actually conscious, potentially conscious, or what have you), they're still the essential components of some as yet unspecified consciousness-generating system. And it's whatever it is about them that makes them capable of generating consciousness that makes them intrinsically intentional; indeed, as I suggested, it is the only thing that gives substance to the intrinsic/derived distinction in the first place.

Now it may be the case (and I don't care, really) that there first evolved (or that one could construct) creatures that had qualia, but no beliefs or desires. They, as in Emily Dickinson's (?) stricture about poemhood, could not MEAN, they could only BE. In that case, the question of what it is that gives a conscious creature real intentional states would have some content of its own. But notice that it would be content that was entirely dependent on whatever it is that makes the creature conscious in the first place. For the minute you try to subtract the capacity for consciousness from the system, you are left with a difference that makes no difference: A creature with "real" (intrinsic) intentionality versus a (turing indistinguishable) creature with mere "as-if" (derived) intentionality.

> The other statement of your claim further down in your message was just
> that anyone with genuine intrinsic intentionality can't be a zombie
> with no possibility of any phenomenal states, now or ever. Surely you
> don't think John disagrees with that?

It’s not that he disagrees with that. But he does not yet seem to be prepared to see the problem of (a) how intrinsic intentionality is realized as (a subproblem of) the problem of (b) how conscious states are realized, or to agree that the question (i) "does a system REALLY-mean P?" is really just the two questions (ii) "Does the system AS-IF-mean P?" plus (iii) "Is there somebody home for whom there is something it is like to mean P?" (irrespective of whether that occupant is actually conscious of meaning P).

The difference between our two respective positions may seem small, but it makes for some big differences in the conclusions Searle would endorse. I, for one, find the "consciously desiring that P" and "unconsciously desiring that P" distinction too vague to base coherent conclusions on. I don’t think the conscious/potentially-conscious distinction can be made rigorous. On the other hand, there are interesting substantive questions too: (1) Can there be conscious states that are not intentional states (as touched on above), and what can be said about them? (2) What is the difference between real intentionality IN A CONSCIOUS AGENT (as in when I’m really understanding Chinese, really doing arithmetic, really getting a drink of water because I’m thirsty, and, perhaps, really referring to John Searle -- I tread lightly in philosophers’ opaque waters here...) and as-if intentionality IN A CONSCIOUS AGENT (like Searle in the Chinese room, my acalculic nephew doing arithmetic as mindless symbol manipulation, a fortuitous outcome in the desert when I’m thirsty, but following what I take to be the directions for finding gold, whereas they are actually the ones for finding water; and finally, with trepidation, when I make a de dicto reference to Searle -- say, as the one who disgarees with P -- not intending him de re, whereas I have nevertheless, unbeknownst to me, picked him out).

> I’m not sure why you say that intrinsic intentionality is redundant
> over and above (phenomenal) consciousness itself as a mark of the
> mental. Unconscious intrinsically intentional states are perhaps not
> always ACTUALLY (phenomenally) conscious--though if John is right, they
> are potentially phenomenally conscious. If this is the heart of your
> disagreement with John, then you DO seem to be insisting that
> unconscious desires always have actual and not merely potential
> phenomenal consciousness. I consider this a possibility, but I can’t
> imagine that this is the disagreeement between you and John. Best Ned

As I said, I don’t think this conscious/unconscious desire/belief distinction is very rigorous. Moreover, I DO hold that either there is something it’s (actually, consciously) like to consciously believe/desire X or there is something it’s (actually, consciously) like to unconsciously believe/desire X. But the substantive point is not the actuality or potentiality of some particular mental state, but the fact that the necessary condition for real intentionality is the capacity for consciousness. No system that lacks the latter has the former, and its capacity for the former draws largely (and perhaps wholly) on its capacity for the latter.

Why is consciousness not a sufficient condition for real intentionality too then, which would make the relation one of it equivalence, thereby doing away with any pretence of independence? Because of the two (relatively minor) unanswered questions I raised: (1) The possibility of consciousness without beliefs or desires, and (2) the possibility of as-if intentionality even in a conscious system (like Searle in the Chinese room). I do consider these questions minor, however. And certainly, as the "mark of the mental," real intentionality, for which consciousness is a necessary condition, is doing no independent work. Hence it is redundant. To suggest that it deserves to be called "the mark of the mental" even in virtue of the two questions I’ve agreed are still open is like suggesting that language is the mark of the mental (false), that awareness-of-being-aware is the mark of the menatl (false), or that the mark of the mental resides in the distinction between two MENTAL states (e.g., consciously doing real math and consciously doing formal symbol manipulation)

Best wishes, Stevan

Date: Wed, 8 Aug 90 01:34:14 EDT From: Stevan Harnad To: dimitri@yktvmt.bitnet Subject: Penrose/Bhaskar/Campbell Status: RO

Dima: I don't find much substance in this, and the little I do find seems very weak. You may forward my comments to whoever you wish.

Styopa

> Date: Tue, 7 Aug 90 18:21:52 EDT
> From: Dimitri Kanevsky
>
> !AUTH 106342! 7 Aug 1990 15:32:54 (#11930) R. Bhaskar
>
> A Pendant to Penrose -- 8/20/90
>
> Submitted by R. Bhaskar, 784-7839, BHASKAR at YKTVMH2
> date: 20 Aug 1990 time: 10:30 place: H1-EF53 speaker: Dr. Robert L. Campbell
>
> A Pendant to Penrose:
> The Interactivist Alternative to Standard AI
>
> The recent visit by Stevan Harnad, and the discussion session on
> Roger Penrose, have brought questions about the foundations of
> AI to the forefront. I will present a brief argument that
> interactivism offers a fundamental alternative to AI, and that
> it offers critical arguments rather different than Penrose's.
> Interactivism presents arguments against both weak AI and strong
> AI, for any conception of AI that presupposes the existence of
> foundational encodings.
>
> Claims:
>
> Weak AI--Turing machine computability isn't enough, because knowing
> the world requires dynamic temporal coordination, and Turing machine
> theory reduces time to a sequence of steps.

Weak AI -- the notion that all systems, including the brain, are computer-simulable -- does not imply that all systems are sequential or even discrete, just that they are simulable sequentially and discretely byt a computer that is turing equivalent to the system being simulated.

> Moreover, interactivism defines consciousness functionally.
> That means that there are accomplishments impossible to a system
> without consciousness. Consciousness is not an epiphenomenal experience
> that humans have but that other systems that behave exactly like us
> could lack. In consequence, a system built according to the precepts
> of standard AI will not be able to pass a stringent version of the
> Turing test.

Consciousness is not something to be defined, and certainly not before there is a highly successful model for it, with lots of predictive and performance power (there is no such model). Hence nothing can follow from a "definition" of consciousness as to what accomplishments are impossible without it. Consciousness IS experience; epiphenomenalism concerns whether or not it plays an independent causal role AS experience. Whether or not other systems are conscious has nothing to do with whether or not consciousness is epiphenomenal. None of these considerations has any implications for whether or not standard AI (pure symbol manipulation) can pass the Turing Test. "Stringent" is undefined; there are two possibilties: Turing indistinguishability in linguistic (symbolic) performance only (the TT) and Turing indistinguishability in Total performance, linguistic and robotic (the TTT). Apart from neural indistinguishability (the TTTT), that's as "stringent" as one can get. Whether or not standard AI (pure symbol manipulation) can pass the TT is unknown, but it is already obvious it cannot pass the TTT because trandsuction is not symbol-manipulation. This has nothing at all to do with consciousness.

> Strong AI--Contemporary AI lacks a tenable conception of representation.
> It presupposes the existence of foundational encodings (representations
> which represent by structural correspondence to the world, and which
> do not stand in for any other kind of representation). Encodingism
> can be shown to be an incoherent conception of representation.
> Interactive representation is an alternative that avoids the
> incoherence of foundational encodings.

No idea from this what "interactive representation" might be, or what the missing "stand in" representations might be, but I'll bet "interactivism" just turns out to be another variant of standard AI, with a few more layers of interpretation to fancy it up...

I don't know what "structural correspondence" means, but standard AI represents by computational (syntactic) "correspondence" only: The symbols and symbol manipulations are systematically interpretable as standing for objects and states of affairs in the world, for example.

> Interesting consequences:
>
> 1. Unlike some objections to standard AI, interactivism isn't
> anthropocentric. It does not deny knowledge to planarians, nor
> to mechanical devices of the right sort. It is, in fact, a form
> of functionalism, albeit a nonstandard one.

"Knowledge" is not interactivism's to affirm or deny to anyone. If the term refers to the usual kind of knowledge, the kind conscious people have, then either planaria and certain mechanical devices do have it or they don't. Who's to say? That's the other-minds problem, and neither definitions nor performance models solve it.

> 2. Interactivism offers a fresh perspective on the metamathematician
> problem, using a hierarchy of knowers, each of which knows properties
> of the knower below it, to deal with reflective knowledge and major
> stages of cognitive development.

>From what little substantive content there is in this summary, I infer that all "interactivism" has done is to help itself to "knowledge" and "consciousness" to describe some form of symbol structure. You get out of that exactly what you put into it.

Stevan Harnad

> Date: 11 Jun 1990 14:55:40 GMT
> From: russell%MINSTER.YORK.AC.UK@CORNELLC.CIT.CORNELL.EDU
>
> Dear Stevan,
>
> I have read your paper on the symbol grounding problem, which
> incidentally I think is a key paper in the development of many new
> concepts in AI and computational intelligence.
>
> However, we have points of disagreement, especially in the area of
> connectionist systems performing symbolic modelling. I am about to
> submit a paper (co-authored with Tom Jackson also at York) in which we
> discuss our thoughts. They can be concisely summarised as stating that
> a hierarchical associative network doing bottom-up feature extraction
> etc. guided by top-down expectations can do symbolic manipulation. This
> is true only if the modular associative units feed knowledge frames (in
> the AI sense)---if this is the case, the knowledge frames can be used
> as the symbol strings, and can be combined in ways explicitly
> formulated in other knowledge fields. Whilst features of the proposed
> system are lost in the summarisation, I hope that you can see the
> principle.

I don't know what "knowledge frames" are, but if they are just symbols, then they are subject to the symbol grounding problem. Moreover, feature extraction is only feature extraction if it's REAL feature extraction from a real sensory projection, not just symbols that are interpretable as sensory input and features.

> The actual reason for me contacting you is to obtain firstly your
> initial reaction to that summary, and to glance at the following
> section of text in which we discuss your paper, to make sure that we
> haven't misrepresented you. Any comments that you have would be

> gratefully received (and, dare I be so rude, as soon as possible would
> help!).
>
> Text follows
> ____
>
> Harnad (reference) utilises similar ideas in his paper on the symbol
> grounding problem. He argues that reasoning can only take place if the
> symbols that are used can be obtained from representations of the real
> world. He proposes that atomic symbols can be combined into composite
> strings that then have a semantic interpretation. He proposes a
> definition of a symbol system that we shall adopt (reference or
> include); whilst all the properties are necessary components of a
> formal definition of a symbol system, the essence is that basic tokens
> are manipulated and combined by explicit rules. These rules are also
> strings of tokens. The manipulation is not based on the “meaning” of
> the symbols; rather, it is done on the “shape” of the tokens, but the
> entire system can be systematically assigned a meaning. Harnad argues
> convincingly for explicit symbolic rules as an essential part of the
> symbolic model of mind, and symbolists believe this model accurately
> represents mental phenomena such as emotion and thought.

The definition was of a symbol system, not a mind; in my view, a pure symbol system (implemented as a physical system, of course) cannot have mental states -- cannot think, reason, mean, in the sense we do -- because of the symbol grounding problem. To ground symbols it is not sufficient to have a "representation" of the real world, for the representation might itself be just symbolic, hence ungrounded. Symbols must be grounded in the nonsymbolic structures and processes that give the system the ability to pick out the objects and states of affairs to which its symbols refer on the basis of the projections of those objects and states of affairs on the system’s sensory surfaces. (By the way, the explicitness of the rules is a contested feature; the essential features are systematic interpretability, of the system as a whole, and its parts and their interrelations, plus the manipulation of symbol tokens purely on the basis of their (arbitrary) shape.)

> Our opinions diverge from Harnad’s when he considers connectionist
> systems, however. He argues that connectionism has no explicit rule
> representation and so cannot act as a model for the mind; instead, he
> takes the stance that connectionism is a dynamical system that relies
> on patterns of activity in a multilayered network, and so cannot be
> viewed as a symbolic system.

I gave explicit criteria for what counts as a symbol system, and it is not at all clear that nets can meet them; the point is not mine, particularly. See Fodor and McLaughlin in the latest issue of Cognition. The systematic, decomposable relations among the symbol tokens appears to be lacking in nets.

> He recognises the power and achievements
> of the connectionist approach in the fields of pattern recognition, but
> finds it difficult to accept as a model for cognition.

I give reasons why it is a mistake to for either nets or symbols to vie for hegemony in cognition. Nets lack systematicity, hence are not symbol systems; and symbol systems are ungrounded. Nets, however, can be used (together with transducers and other analog structures and functions) to ground symbols in a hybrid nonsymbolic/symbolic system.

> He finds the
> connectionist approach suitable for detecting the innnate features that
> are need to ground the symbols in reality, but doesn't accept that the
> methodology is able to do symbolic modelling.

Nets are particularly suited for LEARNED features, not innate ones.

> Much of the arguement of
> this paper is directed towards the realisation that symbolic rules CAN
> be themselves be represented as a pattern of activity, and that the use
> of composite information in knowledge fields are similar to the
> composite strings of the Harnad model; on the basis of these strings,
> the inputs are then manipulated. We feel that a hierarchical
> associative system that uses explicit knowledge fields and hence a
> rule-based approach is able to perform in a way that refutes his
> arguement that connectionist networks are unable to fulfill the
> requirements of compositiveness and systematicity.

I would like to know more about what you mean above, and what kind of a system you are actually describing.

> He states
>
> "nets fail to meet the compositness and systematicity criteria. The
> patterns of interconnections do not decompose, combine and recombine
> according to a formal syntax that can be given a systematic semantic
> interpretation. Instead, nets seem
> to do what they do NONsymbolically."
>
> Our argument is that we can do the symbolic manipulation by
> representing the rules as patterns within the system, and that by
> ordering the architecture in a hierarchical manner, the intermediate
> results that we obtain are congruent to the composite symbols in a
> symbolic system. We would accept that ascribing systematic meaning to
> the ebb and flow of activity in the elements of the network is not
> possible, but essentially this is an implementation issue, in much the
> same way that the firing of a single neuron does not have any symbolic
> meaning, whereas there exists such meaning at a higher level.

I'm afraid I don't see this at all; either there is systematicity or there is not; if there is, there must be explicit symbol tokens, combining in systematic ways. If not, then the symbolic interpretations are not systematically sustained by the system. See Fodor on representations of cups of coffee and "John loves Mary."

> As Harnad
> himself states, being symbolic is a property of a system, not a chunk
> of one.
>

Yes, isolated chunks are not symbolic, nor are they systematic; systematicity comes from the interactions among the symbol tokens. But just as it is not sufficient to ground a symbol system that there exist an interpretation, be it ever so systematic and unique, so it is not sufficient to make a system a symbol system that there exist a holistic interpretation: Systematicity is a property of the interactions of the components of a symbol system. If a symbol system contains the symbol string "The cat is on the mat," parts of it must be the symbol for "cat" and the symbol for "mat," and these must be able to enter into other systematic combinations, such as "the mat is on the cat," "the cat is not on the mat," "the bird is on the mat," etc. Suppose the interpretation is instead "holistic," with the "whole" system representing all of these things, each in a different dynamic configuration; then unless a formal decomposition can be demonstrated that is isomorphic with explicit symbol tokening, there is no basis for concluding that such a system has the property of systematicity in the formal sense discussed here.

Stevan Harnad

> Date: Fri, 8 Jun 90 14:16 GMT
> From: BARRY MCMULLIN <75008378%dcu.ie@pucc>
>
> Thanks for your reply. It clarified matters considerably for me.
>
> You say:
>
> > You misunderstand my grounding proposal if you think it's just
> > imputing special powers to nets that symbol systems lack. I am
> > suggesting that there are ways to put together a hybrid system, made out
> > of transducers and analog transformers, [...]
> > and that such a system could ground a
> > symbol system that uses the names of sensory categories as its (grounded)
> > atomic terms.
>
> And elsewhere:
>
> > [...] it does not much matter to me whether or not
> > nets are symbolic.
>
> Yes, I see now that I was quite mistaken in my interpretation of
> your position, and that, as a result, much of my comment was
> irrelevant. However, there is still a substantive (to me) issue,
> which I'd like to pursue.
>
> First let's kill the background issue: you're convinced (of
> something) by the CR (and related) arguments and therefore see

> an *in principle* problem (SGP); I’m not convinced (of anything)
> by the CR (and related) arguments, and therefore don’t see any
> *in principle* problem (SGP). We’d both be wasting our
> bandwidth debating the CR (and related) arguments, and that is
> not the central issue for this discussion group anyway.

But Searle’s Chinese Room is just one manifestation of the Symbol Grounding Problem. There are others; the Chinese/Chinese dictionary-go-round I described, in which you go on and on looking for the meaning of the symbols from definition to definition, without ever encountering anything but endless , meaningless symbols -- unless you already know Chinese: This is a problem faced by any purely symbolic model of cognition. No CR Argument is needed to understand that.

> The substantive issue to me, which I evidently failed completely
> to explain, is that I think your proposed, practical, attack on
> symbol grounding is seriously mistaken. I think this because it
> looks to me like an attempt to instantiate what Popper has
> (somewhat uncharitably) labled the ‘‘bucket’’ theory of knowledge
> (mind). What I sought to do in my previous message was to
> paraphase Popper’s arguments, in order to demonstrate this
> connection: I now see that was a mistake for I obviously didn’t
> do justice to them.
>
> Now maybe I’m just rehashing issues that were disposed of before
> I joined the list. I am conscious of a comment you made
> elsewhere that:
>
> > [...] I’m always being told that this or that is a philosophical
> > antecedent [...] but usually the similarity turns out to break
> > down or to become trivial.
>
> Nonetheless, I’d still appreciate a comment, however brief: are
> you familiar with the ‘‘bucket’’ versus ‘‘searchlight’’ theories
> of knowledge (and, more generally, Popper’s whole evolutionary
> epistomology), and, if so, what is your response to my claim that
> your symbol grounding architecture is an attempt to realise the
> ‘‘bucket’’ theory?
>
> Regards, Barry.

I have no idea what the "bucket" theory is, and I am unaware that Popper ever worried about how to get robots to do what people can do; however, I suspect that it will turn out to be the problem of "vanishing intersections" that you have in mind: that sensory grounding is neither necessary nor sufficient for thinking because if we look for what is common in the sensory instances of a concept (assuming there even exist any) we find that it is nothing.

I think this argument is false and based on never having even tried, even in principle, to do the robotics. If the "bucket" is something else, you’ll have to say what it is.

Stevan Harnad

> Date: Wed, 13 Jun 90 10:27 MST
> From: MALONEYC@rvax.ccit.arizona.edu
> Subject: Symbol Grounding Problem
>
> Steve,
> I very much enjoyed the symbol grounding symposium at the SPP and would
> be happy to latch onto the skywriting material you mentioned in your
> email message. By surface mail I will be sending you a copy of my
> recent book where, in Chapter 5, I argue on the side of Searle. I was
> quite surprised when, after the SPP session, Eric Dietrich, told me
> that he was less confident in his view of the Systems Solution after
> thinking about my remark to the effect that the Systems Solution
> incorrectly entails that if the person manipulating the Chinese marks
> does not understand Chinese (but rather than some supersystem
> containing the person does), then the same must (but absurdly) be said
> of the person who masters the rules of chess and, thereby, plays chess.
> Well, anyway, the session was worthwhile and rewarding. Best regards,
> Chris Maloney.

Chris, I liked all your contributions to the workshop, as well as your questions from the floor. Did you get to talk to David Hilbert (assuming he was there -- he said he was coming, but I don't know what he looks like)?

I will add your name to the SG list; you may want to retrieve symbol.Z from .princeton.edu in directory /ftp/pub/harnad using anonymous ftp and binary file transfer mode (then uncompressing the file). A partial archive is also there as sg.archive.Z

Your point about chess is on the right track, but not quite on, for the following reason: The correct kind of example to make the point I think you agree with is, say, the hexadecimal arithmetic example, where we teach someone to manipulate hexadecimal symbols (without telling him what he's doing). In principle he could do it proficiently without understanding that he's doing arithmetic, and the "systems reply" would then dictate that HE may not understand, but "the system" still does. This of course is absurd, and Searle's point exactly, in hexadecimal instead of Chinese.

The trouble with your example of learning the real rules of real chess (as opposed to learning a chess-playing computer program in some code whose meaning is unknown to you, which WOULD be equivalent to the Chinese/hexadecimal case), is that the real rules of real chess are all there is to knowing chess (apart from what we know about games, winning, losing, strategy, knights, kings, etc. over and above their chess-specific meanings). The "meaning" of chess (and to an extent that's an odd question, like the "meaning" of tic-tac-toe), apart from the real-world analogies of the game and its components, is just the "meaning" of any game of skill: To know them, you have to be the kind of system that knows a lot else about the world, including what kings and queens are. That includes a lot of robotic knowledge that no symbol-cruncher -- and especially not a mere chess-playing module -- can have, in principle, because it must draw on nonsymbolic structures and processes.

But to the extent that games of skill are really like the Chinese Room to a real person, they are simply beside the point when it comes to the question of whether pure symbol crunchers understand, or have minds. To put it another way, whetever we can and do do mindlessly is not going to be able to help us decide what kind of system has a mind. Things with minds may be able to do things mindlessly, but so what? It's what it takes to be a thing with a mind in the first place that is at issue.

Best wishes, Stevan

Date: Wed, 15 Aug 90 20:02:56 EDT From: Stevan Harnad Subject: Comments on Concept Formation paper Status: RO

Dear Stefan,

I like your Concept Formation paper. Here are some comments:

(1) It is not at all clear what is the optimal (or realistic) proportion of innateness vs. learning in feature detection. I have focused more on learning (partly because I am a psychologist, but partly because even innate feature-detection had to be "learned" somehow by evolution, and so the origins of feature-detection are analogous in both cases).

(2) I am not sure you have fully understood the notion of "context" defined in the CP chapter. The point is rather critical: An invariant feature CANNOT be an absolute one. It can only be a feature that is sufficient to reliably sort members from nonmembers of a category BASED ON WHAT HAS BEEN SAMPLED SO FAR. This provisional sample of confusable alternatives is the context, and what it does is provide constraints on which aspects of the input variation are informative in successful categorization.

Now of course the result of a good, representative sample by evolution may be well-tuned feature detectors that are capable of resolving all the confusable alternatives the species ever has or ever will encounter. But it's still context-dependent in the sense I have described. And in learned ontogenetic categories, as opposed to evolved, innate, phylogenetic ones, this context-relativity can become an important factor.

Now nothing in your paper hinges critically on this point, but the top of page 11 does seem to be bypassing the real point: Of course our senses get continuous, dynamic input, but that does not make it any less context-relative in the sense described here.

(3) I should also point out that whereas the 1987 chapter you cite was indeed very vague about possible feature-learning mechanisms, the Symbol Grounding paper made a more specific hypothesis about connectionism, and subsequent neural network simulations of categorical perception by me and my collaborators at Siemens have shown that neural nets (back-prop) may be good candidates for being both the feature learners and the "carriers" of the CP effect, which seems to be a very robust side-effect of category learning in nets. (By the way, will there be any Siemens people there during my stay in Sankt Augustin?)

(4) Finally, the correct full citation for the Symbol Grounding paper, which has now appeared, is the following:

Harnad, S. (1990) The Symbol Grounding Problem. Physica D, vol. 42, 335-346.

Congratulations on the paper, best wishes, and looking forward to seeing you in September,

Stevan Harnad

Date: Thu, 16 Aug 90 02:33:08 edt From: "Peter Cariani" Subject: Implications of the Symbols & Dynamics view

There are several points that could be added to Dr. Pickering's very thoughtful paper that might heighten the relevance of the ideas for those concerned with symbol grounding as well as the connectionist community at large. Could you pass it on to the Symbol Grounding list? Thanks. Peter Cariani

1) The analogue of the biological evolution of sensors and effectors for a connectionist adaptive computational device would entail the adaptive construction of sensors and effectors. The sensors and effectors of an adaptive device form the basic "semantic" categories through which its computational states are connected to the world.

The problem being addressed here is the problem of finding the appropriate feature primitives (the right sensors) and the appropriate controllers (the right effectors) for a given task. This problem has been almost completely ignored in the connectionist literature (references, anyone?), but it is at least as important as the problem of optimizing the computations performed once one has constructed sensors and effectors adequate to the task at hand. The closest the connectionists get is to say that macro- features can be built up out of logical combinations of micro-primitives, without addressing the problem of how the primitives themselves are chosen. (In the older jargon the former are digital-to-digital operations while the latter are analog-to-digital). Unfortunately, much of the symbol grounding discussion has also taken this tack. It's refreshing to see a paper which doesn't fall into this trap. It's also good to see that Pattee's profound ideas regarding the origins of symbols are beginning to find their way into the discourse. They deserve much more attention within the AI/CogSci/Philosophy community than they have gotten thus far.

2) A connectionist computational network can only be as good as its sensors and effectors. No amount of computation can improve the appropriateness of the sensors and effectors. One cannot do without contingent sensing and effecting elements and still interact with the world at large.

3) Judicious choice of sensors can transform a previously intractable feature space into one which is easily partitionable. A complementary strategy to optimization of computations is optimization of sensors and effectors. Because computational hardware is relatively cheaper (and more flexible) than current off-the-shelf sensor & effector technologies, we tend to think in terms of fixed, relatively unsophisticated sensors. It could be argued that in biological systems it's generally easier to build flexible, more task-specific sensors than it is to build universal computational symbol-manipulating systems. We see more of the former as we go down phylogenetically.

4) Addition of another independent sensor or effector increases the dimensionality of the feature or control space. This is a general hill-climbing strategy: if you're in a local maximum and can't get out, increase the dimensionality of your space (add a sensor or effector). It can only help, and if you're not terribly unlucky the new dimension will provide a gradient for further ascent.

5) All contemporary connectionist devices can be classified as adaptive computational devices (or "syntactically adaptive" devices). I have proposed a class of devices which adaptively constructs its own sensors and effectors. This class of "semantically adaptive" devices is complementary to the adaptive computational devices we have now. In effect it would be possible to automate the process of selecting the sensors and effectors within which a neural net would perform its adaptive coordination. Combining the two kinds of devices would yield a device which not only optimized its syntactic input-output function, but which also optimized the semantic categories within which that function operates.

6) The British cyberneticist Gordon Pask built such a device in the late 1950's. The electrochemical device consisted of an array of electrodes immersed in a ferrous sulphate/sulphuric acid solution. By passing current through the electrodes, iron threads are precipi- tated and dynamically stabilized. The threads have varying resistances, and the network could be trained and rewarded by passing more/less current and thereby altering conductances (weights) between nodes. More interestingly, in about half a day Pask was able to train the network to evolve thread structures which became increasingly sensitive to sound. Following training, the network could discriminate between 2 frequencies of sound, the de novo "evolution of an ear". Yes, the device was rudimentary and couldn't possibly compete with off-the-shelf microphones, but that wasn't the point. The device serves as an existence proof that such a thing is possible.

7) The Pask example is crucial for devising neural networks which increase the dimensionality of their signal space. Networks of fixed computational elements (e.g. McCulloch-Pitts neurons) produce fixed computations. Networks of adaptive computational elements (e.g. ADALINES) yield networks with adaptive input-output functions. All of the behavior of both types of networks can be represented in a model of fixed dimension (once one has all of the input and output state sets of all of the elements and the set of all possible connections, then everything which occurs will be describable as a state-determined (in Ashby's sense) system within that global state set). Networks of semantically-adaptive elements (like the Pask device) can evolve new signalling channels which are orthogonal to pre-existing ones. If it is possible to generate new signal types, then it is no longer necessary to alter many other pre-existing weightings when one wants to make a new association. In effect one is adding new independent dimensions to the signal space, ones which are not logically reducible to those that are already there. (Multiplexing is thus made possible.) An observer modelling the Pask device would suddenly need another observable (i.e. sensitivity to sound) to track the behavior of the device as the device became sensitive to sounds in its environment and began to act contingently upon those sounds. Hence the apparent dimensionality of the Pask device would increase over time. A network of such elements would be constantly proliferating new signal types as well as adjusting the relative weightings between pre-existing signals.

8) I believe that there are plausible biological substrates for such multiplexing/signal generation in the nervous system, involving the tuning of membrane temporal dynamics through adaptive spatial re-allocation of the various ion channel types in axon arbors. Essentially, the phenomena of conduction blocks, calcium oscillations, and tunable electrical resonances can support a neural network paradigm based on pulse oscillators with intrinsic temporal dynamics (e.g. Chung,

Raymond, & Lettvin, 1970; Greene, 1962). Suddenly the holophone memory model of Longuet-Higgins (1970) looks much more biologically plausible, since one has on hand the requisite cellular oscillators and resonant filters. Obviously, there is a good deal of work needed to flesh all of this out. Are there people out there thinking along similar lines?

Some of these ideas will be discussed in the session

Beyond Pure Computation: Broader Implications of the Percept-Representation-Action Triad

at the upcoming IEEE Symposium on Intelligent Control, Philadelphia, September 5-7, Penn Tower Hotel (near UPenn). For information, contact Dr. Herath (215) 895-6758 or Dr. Gray (215) 895-6762.

Considerations of adaptivity, rhythmicity, and self-organization, in both the computational and noncomputational portions of the triad lead to a re-examination of the relations between the computational models of organisms and devices and the world at large.

1) O. Selfridge on the limitations of the current approaches

2) P. Cariani "Adaptive Connection to the World Through Self-Organizing Sensors and Effectors"

3) P. Kugler "Self-Organization of a Percept-Action Interface"

4) G. Pratt "The Role of Rhythm in Neural Informational Processing"

5) P. Greene "A Body-Based Model of the World"

6) J. Pustejovsky "Perceptually-Based Semantics: The Construction of Meaning in Artificial Devices"

References

Cariani, Peter (1989) On the Design of Devices with Emergent Semantic Functions. Department of Systems Science, SUNY-Binghamton, University Microfilms #8926275, Ann Arbor, MI. (1989) Adaptivity, emergence, and machine-environment dependencies. Proc, 33rd Ann Mtg, Intl Society for Systems Sciences (ISSS, formerly ISGSR), Edinburgh, July, 1989, III: 31-37. (1990) Implications from structural evolution: semantic adaptation. Proc. IJCNN-90-Wash I: 47-51. (1990) Adaptive connection to the world through self-organizing sensors and effectors. To appear in proceedings, Fifth IEEE Int. Symp. on Intelligent Control, Phila, Sept 5-7, 1990. Chung, Raymond, & Letvin (1970) Multiple meaning in single visual units. Brain, Behavior & Evolution 3:72-101. Greene, Peter H (1962) On looking for neural networks and "cell assemblies" that underlie behavior. Bulletin of Mathematical Biophysics 24:247-275 & 395-411 (2 papers). There is also a shorter intro, On the Representation of Information by Neural Net Models, in Self-Organizing Systems, 1962, Yovits et al eds, Pergamon Press. Longuet-Higgins (1969) The non-local storage and associative retrieval of spatio-temporal patterns. In: Leibovic, ed. Information Processing in the Nervous System, Springer-Verlag, NY. (Also reprinted in L-H's recent book, MIT press). Pask, Gordon (1958) Organic analogues to the growth of a concept. In: the Mechanization of Thought Processes: Proceedings of a symposium. National Physical Laboratories, November, 1958, HMSO,

London. (1959) The natural history of networks. In: Self-Organizing Systems, Yovits & Cameron, eds, Pergamon Press, New York, 1960. (1961) An Approach to Cybernetics. (intro by McCulloch) Harper & Bros, NY. Raymond, SA & JY Lettvin (1978) Aftereffects of activity in peripheral axons as a clue to nervous coding. In: Waxman, ed. Physiology & Pathobiology of Axons, Raven Press, NY.

Dr. Peter Cariani, 37 Paul Gore St, Boston, MA 02130 email: peterc@chaos.cs.brandeis.edu

----

From harnad Mon Aug 13 23:40:15 1990 To: harnad@learning.siemens.com Subject: Discrete symbols and continuous signals

From: Dr J A Pickering Subject: Grounding Problem

Here, as per my previous email, is the article I mentioned. If you have any comments, corrections, I'd be glad to hear them. Can this go out to the SIG on the Grounding problem? Best wishes, John Pickering.

[The article follows: My only comments would be that it would be important to see the performance power of chaotic attractors actually implemented and tested in robots; at this level it's too hard to say whether they are being used more than metaphorically. Also, learning is probably as important as innate feature detection. Other comments are welcome for readers of the Symbol Grounding Discussion. SH.]

Discrete symbols and continous reality: An ecological perspective on the symbol grounding problem.

(First presented at the 5th. International Conference on Event Perception and Action, Miami University, Oxford, Ohio, July 24 - 28th, 1989. A draft of the complete paper will be available early 1991)

John Pickering Psychology Department Warwick University Coventry, CV4 7AL U.K.

Abstract.

Wigner notes the 'unreasonable' effectiveness of discrete symbol systems in representing continuous dynamics. Likewise Pattee finds it 'incredible' that our only effective representation of dynamics is in formal systems. This paper, extending a suggestion made by Pattee, proposes that the effectiveness of symbolic systems may appear more credible and reasonable if we assume they are not arbitrary creations of human thought but are grounded in the representational mechanisms of natural cognition. Human symbol systems tap into a pre-existent semantics of action which has evolved to reflect the behavioural interchange between an organism and niche.

As Marr and Ullman have pointed out, early sensory information processing does not involve symbolic operation on discrete symbolic elements but nonetheless produces the basis for later symbolic operation. The outcome of this early processing is the detection of significant patterns in the ambient arrays of energies which are caused by the objects and events within the organisms sensory range. The perception-action systems of a given organism are attuned to those objects and events with behavioural significance for that organism and link the resulting activity caused by their detection to appropriate behaviour patterns. This detection is a translation of continuous and rate-dependant spatio-temporal patterns of energy into a discontinuous, rate-independant set of neural states with behavioural significance. Symbol systems may be effective because they reflect this basic and powerful rendering of dynamics already extracted by perceptual mechanisms.

This proposition can be examined in respect of collisions. These, because of their consequences, are particularly significant class of events. Managing contact with the environment is an important function of the perceptual-motor system of all mobile animals. In particular the assessment of impending collisions as beneficial or harmful is a form of cognitive universal. It is highly probable that "smart" perceptual mechanisms in Runeson’s sense make this assessment in relation to an animals behaviour and morphology. Examples are the pursuit-and-capture, prehension, touching and avoidance behaviours of many animals.

A comparative survey of such behaviours shows a phylogenetic continuum from relatively simple taxes and reflexes through to behaviour that is a complex function of transformations in the sensory array, the behavioural and morphological characteristics of the perceiving animal and its momentary purposes. Work in AI emulating behaviour at the more complex end of this continuum relies on elaborate symbolic representations of the environment, the organism (or robot) and the laws of physics. But this work has not been very sucessful in reproducing natural capacities for action. The reason, as Dreyfus notes, is probably the mistaken assumption that to operate in an environment an organism needs a theory of that environment.

Behaviour at a natural grain of organism - environment interchange is not necessarily cognitively mediated. Not, that is, in the sense of conscious strategy or symbolic computation. Nonetheless behaviour of this sort is a phylogenetically fixed pre-cognitive skill rather than a reflex. These skills rely on being able to rapidly detect the behavioural opportunities in the immediate environment and being able to translate what is detected in the ambient array into appropriate action by means of a perception-action semantics. This perception-action semantic wil have evolved to reflect the mutual interaction of organism and niche as have the perceptual and motor sytems of a particular organism. These in turn will reflect the mutually evolved relations between organism and niche.

This perception-action semantics is not a symbolic representational system in the AI sense and cannot be dissociated from the wetware that carries it. A better analogy would be to say that it inheres in the structure of nervous systems in the way that the connectionist approach claims knowlege can inhere in networks. Indeed, the connectionist approach, in some suggestive respects, complements the approach of ecological psychology. For example, ecological psychology’s well known distaste for formalised knowlege and representations as the sine qua non of mental life appears to be shared by connectionism as the following quotations show:"Networks have no representations per se, they merely behave as if they do" and "Networks contain no knowlege, just connections."

Also the strengths and weaknesses of the two approaches are complementary when it comes to the balance between specifiying the constraints on the operations of natural perception - action systems and specifiying the actual neural mechanisms responsible for the operations, as the table below shows:

Ecological Psychology. Connectionism.

Constraint specification: Strong Weak

Mechanism specification: Weak Strong

Recent work on sensory function integrating connectionism with chaos theory suggests how to make the detection via resonance resonance metaphor more specific, as follows: neural networks have a large state space of n dimensions where n is the number of connections in the network. The condition of each connection (its level of excitatory or inhibitory activity) yields a value on each dimension. The state of a network is a position within this state space defined by the condition of each connection at a given moment. The temporal sequence of positions defines a trajectory through this space which in turn defines the behaviour of the network over time. This trajectory will be a function of both the internal structure and dynamics of the network as well as of patterns of activity imposed on the network by energy obtained by the sense organs. Prior to experience, a particular network will have a pattern of connectivity reflecting phylogenetically acquired tuning of the resonant modes that network supports.

During development this a priori connectivity will integrate with connectivity developed over ontogenteic experience. The effect will be to create attractors in the state-space of a network. These attractors will influence the trajectories of a network through its state space. There will be a tendency of the network to relax into the best-fitting or nearest attractor state unless driven elsewhere by the input. This relaxation into attractors translates continuous inputs into discrete states. The trajectory of a network lacking attractors would reflect the external patterns alone and would merely be a continuous response to continuous input. This relaxation may represent a matching by minimisation of the differences between internal constraints and external patterns of excitation acting on the network at any particular time. Indeed, a minimum matching principle might be a mechanism for detecting optimal paths of action with a strong physiognomic and affective character.

However, a priori connectivity is only part of the story and, phylogenetically speaking the mutually evolved relations between organism and niche are handled by a range of mechanisms. These vary from basic neurological perceptual-motor connections in simpler organisms to more elaborate and eventually symbolic systems in more complex ones. As von Foerster has pointed out, this phylogenetic spread is a record of an evolutionary fixation process charting the transition from non-symbolic to symbolic relationships between organism and niche. The unique reflexivity of human cognition has, through a process of cultural fixation, produced formalised symbol systems which externalise the power of this pre-symbolic action semantics. However, it needs to be remembered that the original evolutionary fixation process is the source of the power of formal symbol systems and it is this which grounds the formal symbols of such systems in a pre-symbolic action semantics of natural behaviour.

There are a number of potential empirical follow-ups to this proposition. For example, in respect of impact-management, how rate dependent is impact management under different conditions? How much does reflex avoidance take account of the impacting trajectory? Departures from strict rate dependency will indicate the conditions under which there is internal imposition of discrete responses on continous input.

References:

Skarda, C. & Freeman, W. (1987) How brains make chaos in order to make sense of the world. Behavioural and Brain Sciences, Vol. 10:161 - 195.

For a good discussion of why ’tuning’ is more appropriate here see Cosmides, L. & Tooby, J. (1987) From Evolution to Behaviour. In The Latest on the Best, edited by Dupre, J.. MIT Press.

Maynard-Smith, J. (1987) When learning guides evolution. Nature, 329: 761 - 762.

Date: Thu, 6 Sep 90 23:25:04 EDT From: Stevan Harnad Status: R

Branched to the Symbol Grounding Discussion Group

On Categorization by People and Nets: Is it Easy or is it Hard?

> Date: Thu, 6 Sep 90 22:01:59 -0400
> From: danb@bucasb.bu.edu (Dan Bullock)
> Subject: Re: ART and imposed categorization
>
> Just a note regarding the kind of biasing mentioned by Dan Levine: > [Dan Levine had suggested that categorization problems could be solved by biasing nets toward the relevant features.]
>
> As it happens, I did my dissertation on children’s learning of property
> words. It was jointly inspired by standard Wittgenstein/Russell issues
> (If I call something "red" to a child how does it know that I’m talking
> about an object’s color rather than some other property of the same
> object, e.g. "within a seven mile radius of Fenway Park"?) and by the
> idea that such learning might be explained by a general associative
> learning ability or by associative learning abetted by other
> processes.
>
> The results showed that children were unlikely to guess the correct
> referent property of a novel but consistently used morpheme (like the
> proverbial "wug") if one gave them solely the minimal data needed to do
> so by associative abstraction. Here associative abstraction referred to
> a hypothetical process of associating the word to a series of referent
> objects’ properties until the repeated property (the intended referent)
> emerged from the background on the basis of a stronger associative
> weight between it and the novel morpheme. On the other hand, children
> zoomed right in to make a correct guess if a number of things conspired
> to bias their attention to the referent property. In particular, having
> the property be a value on a task relevant dimension and having the
> value be novel at the time of introduction of the novel morpheme did
> the trick.
>
> Needless to say, given my current occupation, this result did not shake

> my belief in the power of associative learning. However, it led me to
> believe that creatures forced to rely on raw associative learning
> would probably be dead by the time they learned a language. So, I
> think there is much to recommend Dan Levine's emphasis on attentional
> biases. I might add that such attentional effects, as revealed in the
> overshadowing and blocking phenomena of classical conditioning, show
> that most higher animals rarely rely on pure brute force associative
> learning.
>
> On this score, Wittgenstein answered Russell by speaking of biases
> shared by all humans, which biases help get the linguistic community
> off the ground. From there, most hard to form categories get
> transmitted with the help of language, in a myriad of ways. In a former
> life, I wrote a long paper arguing that the march of intelligence has
> been an accretion of devices for getting there faster than is possible
> by brute force association. Most of these devices amount to attentional
> biases, at least in part. Part of my skepticism regarding back prop is
> that it seems like what we might have done if we hadn't become
> intelligent instead.
>
> A final note: after studying the factors that governed infants'
> guesses, we showed that the same factors governed parent's word
> choices. Thus parents generate a biased diet of words that children can
> learn quickly because they share the same biases!
>
> A post final note: By definition, an attentional bias in favor of
> novelty does not beg the question in the way Stevan feared Dan Levine
> was doing by citing his results on featural bias.
>
> -Dan Bullock-

Dan,

Your findings on biasing towards certain features in children's category learning are interesting and relevant to this discussion, but I wonder if they are decisive. I will discuss briefly two related matters. One concerns the question of the true degree of underdetermination of our category learning tasks and skills and the other considers the proverbial example of chicken-sexing, and Irv Biederman's findings about it:

There are two possibilities: Either category learning is hard, and our category learning capacities are prodigious, or it is easy, and our category learning capacities are trivial. (I know there are possibilities in between too, but the poles are relatively easy to describe.) If all the things that people (and animals) ever can and do sort and label have obvious features and boundaries distinguishing them, features and boundaries that are readily found and used, then categorization is easy. If the features are hard to find, and the category boundaries are either hard to find or nonexistent (requiring that they be constructed) -- i.e., if category learning is underdetermined, and hence the problem of feature induction is nontrivial -- then categorization is hard.

If categorization is relatively easy, then of course a precategorical "tilt" toward the winning features will be halpful. But if it is hard, it might not even be known what the winning features are! Consider Biederman's finding on chicken-sexing: The standard story had been that chicken-sexers don't know HOW they do it, they just learn it by trial and error and feedback from masters until at long last they can do it themselves. If an initial "tilt" towards the winning features could be imparted to the novices, it would make their long apprenticeship unnecessary. Yet the masters know no such short cuts. So maybe chicken-sexing is hard.

Enter Irv Biederman, who computer-analyzes photos of chick abdomens, male and female, until he comes up with a winning set of features which he then imparts to the masters who, to their surprise, can now train novices in a much shorter time. So chicken-sexing is easy after all, right? -- But lets not forget the mediation of the computer analysis! The masters certainly had not performed such an analysis, at least not consciously, and they certainly couldn't verbalize and transmit it until they were given the "tilt" by the computer.

What are we to make of this? I think much Wittgensteinian mystification has been spread over the problem of categorization. Some have concluded that categorization is so underdetermined that there does not even EXIST a winning feature set on which to base it (which of course leaves the basis of our success very mysterious indeed). Others have concluded that Wittgenstein's underdetermination problem is solved by simply drawing the category learner's attention to the winning features. This is all well and good when the winning features are known, and when this is indeed the transaction that occurs in category learning. My bet, though, is that in most cases of (sensory) category learning, feature-finding is hard, it is not done consciously, and hence "tilts" are neither available nor used.

In particular, if one-dimensional CP effects are learnable (and I still consider that an "if"), it's no use pointing to the threshold value that must be learned: "See, if it's bigger than that it's an X and if it's smaller than that it's a Y," or something like that. The boundary must be actively formed by the sensory system, from the experience of labeling (and mislabeling). In more complex, multidimensional and underdetermined cases there isn't even a simple threshold or other feature to point to. The confusability among the categories must be resolved by a mechanism capable of finding the invariance that will successfully partition the similarity space in accordance with the feedback on correct and incorrect categorization, and the invariance is not an obvious one.

But I may be wrong, and category learning may be easy, which is really to say that category learning doesn't exist, just a big, negotiable feature-space that can be easily sliced any way we please.

The data will have to decide. To be realistic, we have to see just how hard a categorization task children and adults are actually capable of mastering, and then what kind of model is capable of mastering it, and how. It is cheating to "tilt" the model toward the winning features unless there's strong evidence that we are likewise so "tilted."

Stevan Harnad

Wisconsin Card Sorting vs. Chicken-Sexing

> Date: Thu, 6 Sep 90 23:03 CDT
> From: B344DSL@UTARLG.UTARL.EDU (Dan Levine)
> Subject: Re: ART and imposed categorization
>
> Stevan,
>
> I was not talking about APRIORI biasing toward particular features, but
> deciding what features are important based on external signals.
> Example: in the Wisconsin card sort, as you probably know, the
> experimenter tells the subject "Right" or "Wrong" based on match or
> mismatch in a feature (at times color, at times shape, and at times
> number) but does not tell the subject the reason for the "Right" or
> "Wrong". The subject has to guess why and use it to guide his or her
> subsequent categorizations of later cards. Also, the experimenter at
> one point switches the criterion without warning. (Making the switch
> depends on intact frontal lobes.) So the model builds in some weighting
> of features by "attentional bias nodes" which are in turn influenced by
> positive or negative reinforcement signals. Of course, the model
> sidesteps the question of how the subject decides that color, number,
> and shape are relevant, which may be answered in part by visual
> segmentation models a la Grossberg/Mingolla. But for the type of
> categorization you are interested in, one might be able to assume a few
> relevant features as "atoms", and the problem becomes how to allow
> external signals (from reinforcement nodes or from a knowledge base) to
> modulate selective weighting between those few features. I believe
> Weingard's work and mine are both relevant to that aspect of the
> problem. Best, Dan L.

Dan,

The Wisconsin card sort test is based on simple, stereotyped, overlearned features. No new pattern is learned in the task; just which of the small number of familiar features that are varying is the one the experimenter happens to have picked to be the rewarded one. It is not that such a task has nothing in common with nontrivial category learning, but I'm afraid it's not a great deal. In a nontrivial category learning task the features are many, the right ones are underdetermined and hard to find, and the category of which they are the invariants is the aspect of interest, not just the shuffling and reshuffling of a small number of features. The paradigm task for nontrivial category learning is more like chicken sexing than the Wisconsin card sorting test. It is for this reason that I think feedback-guided attentional shifting from feature to feature may not be a good model for category invariance detection.

Stevan

> Date: Sun, 9 Sep 90 11:43 CDT
> From: Irv Biederman
> Subject: Expert Categorizations
>
> Thanks for your flattering reference to our work on chick sexing.
> However, there is one point in your account that should be corrected. I
> did not do any image processing operations or averaging on chick images by
> computer. Instead, I had the expert circle regions in the pictures so I
> could know where he was looking and then I (myself) looked for a simple,
> viewpoint invariant contrast (which turned out to be convex versus concave
> or flat) in the pictures and in the real chicks as I looked over his
> shoulder when he was actually sexing the chicks. I tested my hypothesis by
> calling out my own classification before he made his.

Irv, I stand corrected. Unfortunately, the part this leaves unexplained is HOW you found the critical features...

> It seems to me that your general account about perceptual learning
> need not necessarily be modified if I was armed with some theoretical ideas
> rather than a computer. The theory (if one could call it that) was that
> difficult (for the novice) shape discriminations, when performed over
> complex and varying images, may be the result of the discovery of a simple
> contrast by the expert at a relatively small scale, IF they are performed
> quickly by the expert. This conjecture was based on the belief that we
> have a single system for object recognition based on simple viewpoint
> invariant contrasts of edges.

You were armed with your geon theory in searching for features in this task. I admire that theory for certain visual shapes, but how will it work for textures, and auditory patterns, etc.?

> On the general issue of symbol labeling, at least for the case of
> entry (or basic) level categories based on shape, I suspect that the
> position outlined in my ’87 Psych Rev paper, which I still hold, would
> place me in the Wittgenstein-Dan Levine-Don Bullock attentional-perceptual
> bias plus linguistics pragmatics camp as to how we ground symbols. The
> adult shares the same perceptual-attentional system as the child and,
> moreover, has the Gricean pragmatic sensibility to know what it is that the
> child does not yet know but would infer--given the child’s perceptual,
> attentional, and pragmatic capabilities--from a given context. My favorite
> example as to how this all works comes from Herb Clark: You are walking
> with a friend and three joggers run by, one of whom is wearing a feather.
> Your friend says, "He’s my neighbor." in the Psych Rev paper I argued that
> the parent is prepared to supply a label for a novel (to the child) geon
> arrangement in response to the child’s asking "what’s that?"

That’s all fine when the game is "I’m thinking of a feature of that object, and you figure out which one it is." And this may well be the game adults play in teaching kids some patterns. But that’s not what conventional chicken-sexers do when they teach their apprentices; it’s not what someone did when you used your geon hypothesis to come up with the concavity feature; and I wonder how close it is to what’s really going on in supervised category learning where the "supervision" does not consist of comunicative interactions between instructor and pupil but environmental consequences from miscategorization (as in eating a toadstool you took for a mushroom on a solitary walk -- shades of the "private language" problem...). The question boils down to: What is really the canonical category learning task? The social reshuffling of a few familiar features, or finding new ones in a haystack?

> But there may not be a single explanation for all categories. I think
> the critical descriptors in your response to Dan Bullock was that the
> categorization that you have in mind was for "one-dimensional, sensory
> categories." It may well be the case that such continua do not yield seams
> which could be exploited by attentional biases. In fact, I would expect
> differences in the operation of the underlying detectors between those
> sensory continua that did (if they exist) or did not allow such reference
> points.

One-dimensional continua are seamless because they are continua. Hence finding the "seam" is worse than hard: You must not just find it, you must "construct" it. Multidimensional continua are even worse. If the features are themselves the (separable) dimensions, then you of course have seams, but if there are many of them -- i.e., if the invariance is highly underdetermined -- then you’re almost in as bad a boat as with a continuum. And if the invariance is disjunctive, conditional, or based on some other compex higher-order rule, you’re back to having to "construct" the seams.

> On another note, we have recently completed a write up of the Hummel-
> Biederman neural net implementation of Recognition-by-Components which has
> the capacity, we believe, to activate a mental symbol from an input of the
> edges representing orientation and depth discontinuities in an object
> image. In the output layer, a unit representing an object’s geons and
> object-centered relations (yes, a grandobject cell [but it could be cells]
> is activated that is invariant to translation, size, and orientation in
> depth. The critical action for solving the binding problem, so that edges
> are appropriately assigned to geons and invariant representations could be
> derived, is accomplished by "fast enabling links" at the input layer that
> synchronize the firing of units that are collinear, parallel, or
> coterminating (to form a vertex). Serendipitously, a derivation of
> attention falls out of the model: Attention is needed to suppress
> activity that would produce accidental synchrony. A copy will be in the
> mails for you. Irv Biederman

Perhaps you’d be interested in describing it to elicit some discussion in this symbol grounding group.

Stevan Harnad

Robots vs. Pen-Pal Modules Again

> Date: Fri, 7 Sep 90 09:31:55 EDT
> From: lammens@cs.Buffalo.EDU (Joe Lammens)
> Subject: About airplanes and simulations. (for SG group)
>
> About simulations.
> A quotation from Stevan Harnad:
>
> [SH: What's at issue is not "replication" in some vague general
> sense, but a specific kind of simulation, namely, symbolic
> simulation, as in a computer or Turing Machine, where the
> syntactic manipulation of symbols on the basis of their shapes
> alone can be given a systematic semantic interpretation. Now
> it seems (according to the Turing/Church thesis, etc.) that
> EVERY (finite, discrete) process can be simulated symbolically
> (and continuous infinite ones can be approximated arbitrarily
> closely). None of that is in doubt. What is in doubt is that every
> symbolic simulation has all the relevant properties of the process
> that it is simulating. Searle showed that this was not so in the
> case of thinking, but we need not go so far, because, as I've
> reminded enthusiasts time and again, symbolic furnaces
> and symbolic airplanes, even though they are semantically
> interpretable as heating and flying, don't really heat or fly. By
> exactly the same token, symbol crunchers that are semantically
> interpretable as thinking don't really think. So much for
> replicating thought -- and that's without even getting to sight (in
> which I argue that thought is grounded). SH.]
>
> JL: There seems to be a lot of confusion about (symbolic) simulations. Of
> course simulated airplanes don't fly, simulated furnaces don't heat and
> simulated rain doesn't get you wet. All of these are simulations of
> physical systems, in a very specific sense to be explained below.
>
> A physical model is a mathematical model, essentially one or more
> equations. The equations describe the behavior of a physical system, as
> measured in a number of parameters. The measurements convert physical
> phenomena into symbolic (numerical) representations, and the equations
> describe the behavior of these measured symbolic parameters. A physical
> model is a good one iff it describes the behavior of the parameters well
> and is able to predict their behavior too (e.g. given the value of some of
> the parameters, predict the value of others, or given the value of all
> parameters at time t1, predict their value at time t2). A model in this
> sense is static: a bunch of equations that can be written down on paper.
>
> One can also use a model to create a dynamic simulation. A simulation is
> essentially a process, not a description. There are different kinds of
> simulations possible. One can simulate one physical system (e.g. a grand
> piano) with another one (e.g. an electric piano). Of course the simulation
> is only partial (an electric piano does not necessarily look like a grand

> piano), as only the relevant properties are being simulated. In the case of
> the piano we consider the sound it makes to be its most relevant property,
> hence that is being simulated. But the electric piano does not produce a
> series of symbols interpretable as sound, it produces sound (although at
> some level in its circuitry it might use a digital or analog symbolic
> representation of sound). Sound is being simulated by sound, or the
> simulator directly replaces the simulated system in its physical behavior.
> There is no intervening level of interpretation, i.e. no systematic
> semantic interpretation is necessary in order to hear the sound the
> electric piano is making. Let's call this kind of simulation a direct
> simulation, or type 1.

Better still, let's call the piano simulation analog simulation; this is not relevant to the symbol grounding problem, which concerns symbolic simulation. No interpretation is needed to mediate analog simulation. Only symbolic simulation is parasitic on interpretation. Nor is "dynamic versus static" the critical variable: Of course we are talking about dynamic implementations of symbol systems here, not their static form on paper. But even a dynamic simulation of an airplane doesn't fly: It's still just symbol-crunching. -- Unless of course it DOES fly, in which case it is grounded (pardon the oxymoron), even if a lot of its innards are just symbolic. The same is true of a symbolic simulation of a piano: As long as it's just symbols interpreted as sound, it's ungrounded. Once real sound is produced (and the system thereby becomes TT-indistinguishable from the output of a real piano of the desired degree of similarity) it's grounded: But at the very least that requires transducers. So a piano could be just a symbol-cruncher plus transducers. A plane is a bit more than that though; a body with a mind surely still more. (So the conclusion that grounding just amounts to sticking transducers onto the right symbol system would not be the correct one.)

> Another kind of simulation is a purely symbolic one, where we create a
> dynamic symbolic system (e.g. a computer program) that models the behavior
> of a physical system over time by producing a series of (potentially)
> varying parameter values (i.e. symbols). An example would be a program that
> produces a series of numbers representing the sounds that a grand piano
> makes when being played. Here we do need a systematic semantic
> interpretation to make sense of the simulation, e.g. interpreting the
> numbers as describing a sound spectrum in time. This is also the class
> where symbolic simulations of airplanes, furnaces and rain belong. Since
> their output consists of symbolic parameter values and not actual physical
> behavior, there is no confusion possible with 'the real thing'. Of course a
> series of numbers doesn't fly or give off heat or get you wet, or sound
> like a grand piano. Let's call this type of simulation indirect, or type 2.
> Note that a type 1 simulation may contain a type 2 simulation, as noted
> above, though that is not necessary.

The presence or absence of transducers and analogs OF THE VERY BEHAVIOR THAT IS BEING SIMULATED is the critical difference between your type 1 and type 2 simulation. The dynamic implementation of a symbol cruncher (type 2) is just a computer program running on a computer. The type 1 dynamic implementation of, say, a piano, has to be more than that, because it needs a way of producing acoustic waves; same with the type 1 implementation of a plane. I'll leave the case of the thinking mind as an exercise. Here's a hint, though: For some kinds of behavior (symbols-in/symbols-out), their type 1 and type 2 implementations would be the same; computing is an example of such a behavior; the standard (purely verbal) Turing Test (TT) is another. For other kinds of behavior, their type 1 and type 2 implementations are necessarily different; music, flying, heating are examples; so is the the (robotic) Total Turing Test (TTT).

> Now, the simulation of intelligence, as defined in the classical Turing
> Test (TT), is clearly not a type 2 simulation. Here we are not modeling a
> physical system through symbolic parameter values. Let's confine ourselves
> to the classical Turing setup for the sake of the argument. In
> (simulatedly) interpreting/manipulating/crunching what someone types to it
> and typing things back, the computer/program/simulation is not being fed a
> symbolic measurement of system parameters and producing other symbolic
> parameter values; it REPLACES the modeled system completely, behaving as it
> would. That is the whole idea behind the TT. It is a direct simulation
> (type 1), i.e. *exhibiting* the relevant physical behavior, as opposed to
> simulations of airplanes that are one step removed from reality in using
> symbolic parameter values that *represent* physical behavior, given the
> right interpretation (type 2). An equivalent simulation of an airplane
> would be something that replaces it and produces the same behavior, i.e.
> something that flies. When the human is replaced by the computer in the TT,
> and the person at the other end of the communication link cannot guess who
> is who, then the simulation is successful. Again, success is not defined
> as producing the right symbolic representations of system behavior, but as
> producing the right BEHAVIOR ITSELF. When we want to compare this to other
> kinds of simulations, we should compare to behavioral (type 1) simulations,
> i.e. success for an airplane simulation would be defined as the simulator
> actually being able to fly, for the simulated furnace to produce heat, for
> the simulated rain to be wet, for the simulated piano to produce sound.
> The behavior in the physical world must be indistinguishable, relative to
> some criterion such as the TT.

But unfortunately people are able to do a lot more than be pen-pals. So it is quite likely that their pen-pal behavioral capacity draws on their other behavioral capacities -- unlikely that it is some sort of isolated, independent module. So successful TT-passing capacity probably requires successful TTT-pasing capacity, and that's type 1, not type 2. The TT is simply equivocal about whether or not it draws upon nonsymbolic (type 1, analog) processes. If it can be passed by symbol-crunching (type 2) alone, then it is open to Searle's objection, which I think is quite valid and decisive. But if (as I suspect) it cannot be passed by symbol-crunching alone, and depends instead on type 1 processes -- is GROUNDED in them, in fact -- then the right test is the TTT and the question becomes moot, because the candidate is hybrid nonsymbolic/symbolic rather than just a type 2 dynamic implementation of a pure symbol cruncher.

> The difference between a symbolic airplane simulator and a symbolic
> intelligence simulator is that the former is a type 2 that models a
> physical system through the intermediary representation of parameter values
> (symbols), while the latter models behavior by behaving. It's a type 1
> simulation with an embedded type 2 simulation. AI programs are not
> simulating brains but intelligent behavior, just as the electric piano does
> not simulate strings and hammers but sound. When you get out of a flight
> simulator session, you're still in the same place (the relevant property is

> physical movement through the air). When you get out of a TT session,
> something has changed: you have talked to a system about something (the
> relevant property is carrying on an intelligent conversation).

A flight simulation may give me vertigo, but that does not mean I've been airborne. There's some confusion arising about who/what is doing the testing and who/what is being tested here: In the TT, the computer is being tested, and the hypothesis was that if it could behave indistinguishably from a pen-pal, then it really understands what it's writing and reading. Searle correctly shows that this cannot be so, because he can implement the same program without understanding. The symbol grounding problem suggests WHY a pure symbol cruncher cannot understand: Because its symbols are not grounded in the objects to which they (can be interpreted to) refer. (Remind yourself that a pure TT-passer would be stumped if asked to point to a cat.)

A candidate that can pass the TTT does not have this problem, but it is no longer a pure symbol cruncher either. Just as a plane can really fly, a TTT-passer has ALL of our robotic capacities in the real world. A mere TT-passer is INTERPRETABLE as engaging in an intelligent conversation, but that interpretation is merely projected onto it by us, exactly as it is in the case of the airplane simulation. One cannot redefine "behavior" in such a way as to make symbol-crunching anything other than what it is: symbol-crunching, be it ever so interpretable as - whatever. The TT, in other words, is just as type 2 as the airplane simulation: Symbols in, symbols out, and the rest merely interpretable as flying and understanding.

It is the confusion about the tester and the testee that has made you interpret the simulated airplane example as a flight simulator that fools you into thinking you're flying. I didn't mean anything as (irrelevantly) tricky as that: Only the legitimate airplane simulation that might go into testing hypotheses about real flight. Computer airplane models, where no one thinks they really fly or wants them to; rather, they are being used to symbolically test models (type 2) before they're actually built into prototype planes (type 1). The power of computer simulation (and Church's Thesis, and Weak AI in the case of mind-modeling) is that this method can indeed test hypotheses about real flight and real thinking by simulating their properties formally, i.e., symbolically. But it's PRECISELY the same mistake to suppose this symbolic activity to be the real thing in the case of pen-pal correspondence as in the case of flight.

> Once this difference is clear we can do away with the confusion about
> flying simulated airplanes, and the whole question of whether a system that
> passes the TT is truly intelligent (truly understands) or not becomes
> rhetorical. It does not matter, and it is undecidable. After all, what
> other evidence do I have to decide that any person I am talking to is
> really understanding and intelligent, and not some improved R2D2 running a
> very good simulation of intelligence? The only evidence is his or her
> behavior. For all I know Stevan Harnad is an AI program, for all I've seen
> of him is written text via e-mail. Yet it is far from me to claim that he
> does not understand what I write, program or not. This is also not meant
> to imply that the Symbol Grounding problem is not a valid one or that we
> could do without symbols being somehow grounded in perception, only that I
> do not think that "simulated airplanes don't fly" is a valid objection
> against symbolic AI.

Ah me! First there is no difference between a real thinker and any TT-passer, and now it doesn't matter -- and doesn't matter because you can't be sure *I* think either. Well there is a difference, and each of us knows perfectly well exactly what the difference is in our own singular case: There's somebody home. Of course, it is perfectly possible that every entity, great and small, part and whole, and every complex permutation and combination thereof, including this keyboard conjoined with alpha centauri, has a mind, but we must not let that crowded possibility make us give up in despair. It is also possible that many things DON'T have minds. First, we must not lose sight of the fact that there IS a difference at issue. Next, we must not insist (arbitrarily) that only differences for which we can have evidence can be real (otherwise there's only ONE real mind: one's own). And finally we must recall with relief that TT-passing is NOT everything we are capable of doing, any more than chess-playing is: The TTT is everything we are capable of doing, and the TTT can only be passed by a grounded (hybrid) system. So ungrounded symbol systems -- even hypothetical TT-passing ones -- are nonstarters from the outset. Now, even a TTT-passing robot (like R2D2) may not have a mind -- so much for what can be known for sure -- but we were only talking about the necessity of groundedness, remember? It's the problem of grounding symbols that's at issue here.

> Finally about Searle's Chinese Room argument: it's irrelevant, in my humble
> opinion. His argument is analogous to claiming that in a computer running a
> Chinese understanding program, it's the CPU (i.e. the thing that interprets
> the program and does the symbol manipulation) that does or is supposed to
> do the understanding. I think no one has ever claimed that about a CPU
> running an AI program, or any other program for that matter. No one seems
> to believe that a CPU really 'understands' lisp (in the sense that we
> understand it) when it's interpreting a lisp program. Yet to the outside
> observer it does, since the machine displays the right behavior. Searle's
> argument is also analogous to claiming that the actual nerve cells in our
> brain (which may be regarded as simple CPUs, just a whole lot of them)
> understand English or are intelligent, which again no one would want to do
> I think. Call this a systems argument if you want; I don't care much for
> pigeon holing.
>
> Joe Lammens.

Unfortunately, I don't just call it the system argument; I also call it wrong. When Searle memorizes all the symbols and rules he is all there is to the system. There's nowhere else to point. And if one has subscribed to a hypothesis whose punchline is that under those conditions Searle himself would no longer be the relevant judge of whether or not he understands Chinese -- that there's someone else home in there, or what have you -- then I think it's about time to abandon that hypothesis.

Stevan Harnad

SIMULATION AND REALITY

> Date: Wed, 19 Sep 1990 07:09-EDT
> From: Hans Moravec CMU Robotics hans.moravec@CS.CMU.EDU
>
> Dear Steve, Your argument is nicely clear, but not entirely convincing.
>
> Let's suppose the Pen-Pal simulator in the TT test actually
> believes itself to be a human being--lets say the whole nervous system
> structure and mental state of some unsuspecting human was recorded at
> some instant then translated into a program, inserted into a computer
> and activated.

Hans, unfortunately, with that one simple assumption, you have simply made anything that follows beg the question. For the question is whether "translating the structure of the nervous system and mental state into a program on a computer" yields something that can have beliefs at all -- i.e., whether it can have a mind. The tentative answer, for a variety of reasons, including Searle's argument as well as the symbol grounding problem, is "No, a program is not enough."

> The questions from the tester must appear to this
> simulated person in some form. Let's suppose the interface is through
> the part of the data structure that emulates the retina, and the
> questions manifest them as glowing letters. Also the (simulated) motor
> and sensory nerves for the fingers are connected to code that simulates
> a keyboard. So the TT simulation remembers a normal human life, and
> suddenly finds itself in front of what appears to be a computer
> terminal running a talk ring in a very, very dark, very, very quiet
> room. The TT simulation may answer a few questions reflexively, but
> will begin to worry about why it can't sense anything but the text, and
> can't remove its hands from the keyboard.

So far, all you have is symbol-crunching that is interpretable as "glowing letters," hand and a keyboard (perhaps even symbol-crunching that would be capable of driving the pertinent screen, effector and keyboard transducers, if they were there to drive). But so what? That's just interpretable symbol-crunching. The question is whether there is anyone else around (except us) to be actually EXPERIENCING what all this code is interpretable as being. One can't simply assume that the answer is yes: There's still Searle's argument and the symbol grounding problem to contend with. The former shows that no one is home under these conditions and the latter suggests the reason why: Because the system cannot actually pick out the objects to which its symbols refer. To reply that the simulation software could do it if it were only hooked up in the right way to the right transducers is not only yet another assumption (which I think is wrong), but even if it's true, it no more indicates that the isolated symbol cruncher has a mind than does the same assertion about a computer running no program at all ("It could do it if it only had the right program AND were hooked up to the right transducers in the right way"). To meet the TTT criterion, everything needed in order to have the behavioral capacity must be there actually, not just potentially.

(Nor will it do to grab the other end of the stick and remind me about blind people, paraplegics, etc., who still have minds. We don't know what ELSE they have that underlies their deafferented and paralyzed behavioral capacities. There's an awful lot of intact nonsymbolic stuff in their nervous systems, whereas the hypothesis on trial here is whether all that's left is a symbol-cruncher.)

> Since the growing panic would spoil the TT (not to mention the
> inability of the simulation to work things out on scratch paper, or to
> stare into space for inspiration), let's surround the glowing letters
> with an ultra high quality, realistic physical simulation of a
> terminal, and that with a whole office, all faithfully interfaced to
> all the I/O portions of our simulated nervous system. Of course to make
> this interface faithful we also need to simulate the physical body of
> our TT simulation. Now the Turing test can continue. If the
> questioner asks "Can you see a cat where you are?", the TT simulation
> may be able to look around and answer "Yes" truthfully, if we had
> thoughtfully included a simulated cat in the simulated room. But if
> the question is "Is it raining there?", and the TT simulation goes to
> the simulated window and looks out, the jig will be up unless we also
> have a good simulation of the outside (with movement, sound, wind,
> smells, other people).

First, there's no panic unless there's someone there to panic, which is the contention that's on trial here. Until further notice, all you are describing is symbol crunching that is amenable to a systematic interpretation (as simulating a terminal, office, brain, body, cat, etc.). Even if the program could actually simulate (some of) these things (I'm doubtful about the brain) FOR US, if attached to the right transducers, what's at issue here is not what it could do for us, but whether there is anybody in there that it's doing it all for right now. And despite all the interpretations projected onto it, the only thing that's really going on in there is symbol-crunching.

True symbol-grounding, by the way, must be grounding to real objects in the real world. A simulated body in a simulated world may allow us to do a lot of hypothesis testing about HOW to successfully implement a real body passing the real TTT in the real world, but it would no more have a mind than a simulated plane could fly -- even though the latter too could be used to do a lot of hypothesis testing about how to successfully implement a real plane in the real world.

The foregoing point is awfully simple, yet it is consistently misunderstood, probably because of the hermeneutic power of projected interpretations, and of freely substituting oneself for the candidate in the simulated situation. Another way to look at it is that it is unlikely that a symbolic simulation could successfully anticipate all the contingencies that a successful TTT-passer could handle: For that you'd need not just a TTT-indistinguishable robot but a TTT-indistinguishable world: A tall order, but perhaps not a logically impossible one. No matter: "E pur NON se muove!" That grand simulation would still not have a mind, any more than it had any of the real properties its world-simulation component was simulating. It would still just be a symbol-cruncher, merely INTERPRETABLE as doing and being all those things. It would, however, contain all the relevant information for implementing the real thing. The real thing, however, would be a TTT-capable robot in a real world; and only that implemented robot would have a mind, just as only the implemented plane in the real world could fly.

> Though this process could probably be stopped at some tolerable
> scale with a simulated backdrop, suppose we continue it to the limit
> and create an entire simulated universe of some size. There is no need
> to ground the resulting grand simulation--it is complete within itself.
> (After all, we have no reason to believe our own physical universe is
> grounded in any way.) So we have an effective Turing Test simulator

> with no external grounding.

One of the sure symptoms of getting lost in the hermeneutic hall of mirrors created by projecting mentalistic interpretations onto otherwise meaningless symbol crunching is that one forgets the distinction between simulation and reality. Hans, there ARE real nonsymbolic objects and processes like tranducers, analog systems, heat, flight. The world is NOT just a big symbol-cruncher that is INTERPRETABLE as all of these things.

> Steve's TTT approach short circuits the necessity of the above
> expansion of the simulation by connecting the TT simulation to our
> actual physical world through a physical robot body. (In fact, the
> boundary between the simulation and the physical could be at other
> stages, for instance the windows of a simulated office could look out
> at physical surroundings through appropriately mounted transducers).

The TTT constraint performs only two functions, but they are both critical: First, it expands the empirical demands of mind-modeling to the degrees of freedom of a normal scientific-engineering theory, Such a theory is still underdetermined (there's more than one way to skin a cat, or even design a physical universe) but not grotesquely so, as a mere toy problem or module would be (one that could only play chess or only describe some scenes, etc.). That's one of the reasons for the "Totality" requirement of the TTT.

The second reason is that it makes the empirical constraints exactly identical to the one we face with one another, in contending with the everyday other-minds problem: We have no stronger grounds than the TTT for assuming anyone else but us has a mind either. Evolution, the blind watchmaker, doesn't have stronger grounds either, though biomolecular constraints may eventually necessitate a "TTTT," which calls for Turing-indistinguishability not only in our total bodily behavioral capacities, but in the behavioral capacities of parts of our bodies too, perhaps right down to the neurons and molecules. Empiricism cannot ask for more. It is of course risky to legislate in advance how much of our TTTT-capacity may be relevant to having a mind, but, as I've written,* I'm fairly confident that all the substantive questions will have already been answered once we can build and understand systems with TTT-capacity, and that the rest will just be the fine tuning.

The candidate's need to have the capacity to pick out the real objects and states of affairs that its symbols refer to in order to ground their interpretations in something other than our own projections is self-evident, it seems to me. Symbols pointing to simulated objects are just hanging from a recursive skyhook rather than being grounded (what in turn grounds the interpretation of the simulated-object-symbols?). Now I have already agreed that simulation can successfully anticipate and second-guess a lot about what it takes to ground a real system (that's Church's Thesis and Weak AI), but that does not make the simulation the same as the real thing. And I have also given reasons why a real TTT-passer will not just consist of simple transducers hooked up to a symbol cruncher that's doing the real work. I'm betting there's more to grounding than that, and that the requisite nonsymbolic structures and processes play a substantive and essential role in it, indeed a more fundamental role than the symbolic component.

> Turing's traditional approach instead has an extra-intelligent
> agent who uses its additional smarts to lie. To make the lies
> consistent, the machine would have to imagine a consistent world about

> which it answers questions. This imagined world is very like the
> universe simulation in my original scenario, though perhaps more
> efficiently encoded and created in a "lazy evaluation" mode where
> things are decided only as needed to answer questions. (But the
> universe simulation of my original scenario could have the same
> economizations (and isn't that the way our physical quantum mechanical
> world works? A situation is in an undecided "mixed state" until a
> measurement forces it to "collapse" to one of the
> alternatives--apparently creating new information out of nothing (or
> from *somewhere*) by forcing a choice)).
>
> Hans Moravec CMU Robotics hans.moravec@CS.CMU.EDU

This imagined-world in the imagined-mind of a pure symbol cruncher sounds like just that: Imagination, a sci-fi scenario. As to the role of quantum mechanics; it seems inadvisable to try to trade off one mystery (the mind/body problem) with another (the quantum paradoxes). Better to try to solve them or let them simmer independently, as they are, to all intents and purposes. But for those who are interested in seeing the quantum gambit given a run for its money, see the forthcoming December 1990 issue of BBS, in which there will be critical reviews of Roger Penrose's "The Emperor's New Mind," from 37 neuroscientists, computer scientists, psychologists, physicists, biologists, and philosophers, together with the author's article-length precis of the book and his Response to the commentators.

Stevan Harnad

Reference:

Harnad, S. (1991) Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem. Minds and Machines 1: forthcoming.

> Date: Thu, 20 Sep 1990 05:02-EDT
> From: Hans.Moravec@ROVER.RI.CMU.EDU
> Subject: Re: Simulated robots/Simulated worlds
>
> > From: Stevan Harnad
>
> > Hans, unfortunately, with that one simple assumption, you have simply
> > made anything that follows beg the question. For the question is
> > whether "translating the structure of the nervous system and mental
> > state into a program on a computer" yields something that can have
> > beliefs at all -- i.e., whether it can have a mind. The tentative
> > answer, for a variety of reasons, including Searle's argument as well
> > as the symbol grounding problem, is "No, a program is not enough."
>
> Oh, I don't care if it's a conventional program. Call it a
> neuron level nervous system simulator, maybe implemented on neuron-like
> hardware. Or extract the victim's actual nervous system from his body

> (with suitable life support) and replace the neurons one by one with
> sufficiently good artificial prostheses. (Of course, I would like to
> hear a rationale for believing such a purely physical system couldn’t
> be simulated with sufficient precision on a sufficiently powerful
> conventional computer.)
>
> >. Hans, there ARE real
> > nonsymbolic objects and processes like tranducers, analog systems,
> > heat, flight. The world is NOT just a big symbol-cruncher that is
> > INTERPRETABLE as all of these things.
>
> I think this is a fundamentally false statement. I see no reason
> to doubt that *From the inside* the real world is not distinguishable
> from a sufficiently good (including sufficiently complex) simulation of
> itself. This seems to be be the crux of our disagreement. I think it
> could be that you and I are in somebody’s simulator right now. You
> think that isn’t possible, that "real" reality is somehow
> distinguishable. I challenge you to propose a scientific experiment
> that could distinguish the two possibilities.
>
> Hans Moravec CMU Robotics hans.moravec@cs.cmu.edu

Hans, I think you are confusing two distinct questions:

(1) Can a system with a mind tell whether it’s in a real world or a simulated world. Answer: No, but so what?

(2) Can a pure symbol cruncher have a mind? Answer -- according to me because of the symbol grounding problem and according to Searle because of the Chinese Room Argument -- No.

It DOES matter for (2), but not for (1), whether the candidate system is just a symbol cruncher or some other kind of system. There ARE scientific ways of telling the difference there. (2) is at issue here, not (1).

Here’s an experimental prediction: Only a system capable of passing the TTT could pass the TT. It would follow from this prediction that since a pure symbol cruncher could not pass the TTT (because of the empirical difference between computers and transducer/effector/analog systems), a pure symbol cruncher could not pass the TT.

That’s a tall order, but testable. Here’s another experimental prediction: A TTT-passer will not be just a symbol cruncher hooked up "in the right way" to transducers and effectors. Nonsymbolic structures and processes will play an essential role in its TTT-passing capacity.

None of these can settle whether or not the TTT-passer has a mind, of course, but the symbol grounding problem concerns necessary conditions for this, not sufficient ones.

Stevan Harnad

To: Symbol Grounding Discussion Group

Here is the description for two symbol grounding projects from Jim Geller at NJIT, with some comments from me. If you have grounding or grounding-related projects underway, please do describe them to the group, particularly if feedback would be helpful.

Best holiday wishes, Stevan Harnad

> Date: Mon, 24 Dec 90 12:36:02 EST
> From: geller@vienna.njit.edu (james geller)
>
> (1) Symbol Grounding. B.P. Sen
>
> In this work a program was designed that reads pairs of sentences and simple
> pictures. The sentences are processed with an Augmented Transition
> Network (ATN) interpreter. The pictures are either prestored or entered
> by a user by typing in a two dimensional pattern such as
>
> 111111111111
> 111111111111
> 111 111
> 111 111
> 111 111
>
> which incidently stands for a table. The sentences may contain one
> unknown word each. The basic assumption of the system is that if there
> is only one unknown object in the image and one unknown word in the
> sentence, then the word describes the sentence.

Interesting, but I would make two points:

(1) As I conceive it, grounding must reside in the capacity to pick out objects from their sensormotor projections. Pictures are not sensory projections. They are themselves just symbol strings (interpretable as pictures). I think this is a problem for your approach, though, as we discussed before, you might be able to find some recursive solutions that simply presuppose bottom-level grounding and examine some of the features of the higher-order combinatorics of grounded symbols and the inheritance of their grounding constraints.

(2) Related to the above, I don't think one unknown name and one unknown picture quite captures the problem of inheriting grounding, because in that situation simply attaching the new name to the new picture would "solve" the problem trivially. There has to be more to it than that.

> The system stores the picture-word association, so that it can later on
> recognize the object if it occurs in another picture, and it can also
> show how the object looks, if it is asked. E.g., if the system is asked
> how a table looks, it will reprint the above picture. In a later sentence
> the newly acquired association may be used, so that one can have

> a sentence like "the vase is on the table", and the system will acquire
> the new (unknown) word vase in this way, assuming there is a picture with
> a table and an unkown object.
> The system limits its search space by "understanding" terms like LEFT OF,...
> and ignoring irrelevant parts of an image.
>
> (2) Symbol Grounding Paradigms - A Language Learning Perspective.
> by A. R. Simha.
>
> This project contains a design and implementation of a system that
> acquires unknown words by using a small set of explanation patterns.
> Again it is assumed that all but one of the words in the sentence
> are known. The system will then judge from the form of the sentence
> what type of grounding is used. We permit bottom-up, top-down, identity,
> and similarity grounding, and small variations of those. The system will
> then assert the information about the grounded object in a semantic network.
> This will be done by storing an assertion of the form "word X is grounded
> by word Y using grounding type Z." If the newly grounded word is used
> in an upcoming sentence it can be used to ground newer words.
>
> The purpose of this project is to understand how children and dictionary users
> ground new symbols with old ones, what you have called "bootstrapping
> from a kernel" (we are citing this of course).

Interesting questions, but it’s not clear whether you are actually answering them (i.e., grounding symbols) or merely doing symbol manipulations that are INTERPRETABLE as grounding. Remember that it was getting from interpretable-as-if X to really X that was the motivation for symbol grounding in the first place (at least for me).

Another way to put this might be: How would your system help someone trying to implement a real robot in the real world? To the extent that it could do so (in a way that could conceivably scale up to the Total Turing Test), your grounding proposal would be real.

> I hope to write this work up in a paper which I will be glad to send you as
> a preprint before submission.

I’d be interested in seeing the paper. May I post this (and further exchanges on the subject) to the symbol grounding list? I think feedback on these important issues would be beneficial to all.

Date: Mon, 24 Dec 90 15:56:56 EST From: geller@vienna.njit.edu (james geller) To: harnad@clarity

Yes, feel free to post it! Jim TO: Symbol Grounding Discussion Group (309 lines)

GROUNDING: REAL VERSUS SIMULATED

> Date: Fri, 11 Jan 91 16:51:34 -0600
> From: uhr@cs.wisc.edu (Leonard Uhr)
>

> Do I understand correctly: you feel that Jim Geller's program, which
> inputs "pictures" (built of zeros and 1's, using an editor) and simple
> word-strings is NOT properly grounding the learned words to
> sensory-motor events? Since this is such a nice simple and concrete
> example, I'd like to try pursuing what WOULD BE adequate grounding:
>
> If the picture had been input by a TV camera?
> Or drawn on a tablet by a person?
> Or the system had a robot cart that wheeled the TV camera toward the table?
>
> and aimed the camera? and focussed it and pushed its on-button?
>
> What if the very simple pictures that have already been used were
> actually the result of input from a TV camera? or a robot cart, etc.?
>
> I'm keeping things simple to stimulate you to object, and MAKE THEM
> ADEQUATE. If none of these is sufficient, what do humans or other
> animals do in addition? Would a human being who manipulated carts and
> robot arms on the moon or in the next room, using TV cameras, etc. no
> longer be grounded?
>
> If this isn't a good example, do you have any ideas for a better one?
> I'm pressing the concrete since theoretical discussions, although very
> interesting, are very hard to pin down.
>
> It seems to me that the sensory/perceptual interface to the "real world
> out there" certainly does introduce an overwhelming richness and
> complexity of information - posing great problems in terms of
> processing and understanding it, and also great potential rewards. I
> like the idea of grounding as needing this great complexity of the real
> world, and also perceptual processes far more complex than
> transduction. But then it becomes a matter of degree-of-complexity and
> there is no sharp line. We could model pieces of the real world inside
> the computer (just as we are already doing when we model a vision or
> language-learning system). Are you arguing for something more?
>
> Len Uhr

Good questions, but readily answerable, I think: What we are trying to ground is a symbol system. Symbol systems (e.g., books, computer programs, computer programs actually being executed on a computer, and, if we are symbol systems, our heads) have the important property that they are semantically interpretable. The symbols and symbol strings can be coherently and systematically interpreted as MEANING something, e.g., as referring to and describing objects, events and states of affairs in the world.

The symbol grounding problem is that in a pure symbol system (e.g., a computer, which only manipulates symbol tokens on the basis of their shapes, which are arbitrary in relation to what they "mean") the meanings of the symbols are not grounded: They are completely parasitic on the meanings we project onto them in interpreting them. That is why the symbol grounding problem is well captured by the idea of trying to learn Chinese from a Chinese/Chinese dictionary alone: All you can do is go from meaningless symbol to meaningless symbol, based on the shapes of the symbols, in endless circles. The dictionary meamings are not grounded; they depend on our already knowing some Chinese (or some mapping of Chinese symbols onto English ones, where we already know English).

We have turned to cognitive robotics as a way out of this symbol/symbol circle: If a robot has the capacity to pick out, from their physical interactions with its transducer/effector surfaces, the objects, events and states of affairs to which its symbols (or a kernel subset of them) refer, then those symbols (and higher-order combinations of them) are grounded in this robotic capacity. It is no longer true that the meanings must be mediated by the mind of an outside interpreter.

If the robot internally tokens "cat" and "mat" in the string "The cat is on the mat" and someone challenges us that the meanings of those symbols are merely being projected onto the robot by outside interpreters then we can point out that the robot not only symbolizes, but acts in the world in a way that is systematically coherent with its symbolization: It can pick out "cats" and "mats" and "cats being on mats."

Now a toy demonstration of this would not be sufficient for real grounding: It must scale up to the Total Turing Test (TTT): The robot’s verbal and sensorimotor performance capacity must be indistinguishable from our own. But neither would a simulated demonstration be sufficient, one in which there was no real robot, no real world, just SYMBOLS that were being interpreted as if they were real objects, real sensorimotor interactions, etc. For that would simply put us right back into that symbol/symbol circle that the grounding was supposed to get us out of!

So this is why ignoring the real world, ignoring real transduction, and simply dubbing a matrix of 0’s and 1’s an "object" or the "sensory projection of an object" will not do. The TTT must be passed in the real world, by a real robot; that’s the only constraint that will logically answer the objection that otherwise all its interpretations are merely ungrounded projections by the minds of outside interpreter.

Now don’t misunderstand me: I’m not saying that a grounded robot may not, somewhere in its innards, make some use of a matrix of 0’s and 1’s to help accomplish its grounding. That’s an empirical question. But the grounding does not consist in the relation between that matrix and other symbols. It consists in the robot’s relation to the world, or, more specifically, whatever internal structures and processes -- beginning with the proximal sensory projection originating from the distal objects, events and states of affairs in the world, and ending with the effector projection and its actions on the objects, events and states of affairs in the world -- and the CAPACITY it gives the robot to perform TTT-indistinguishably from us.

I am not saying a 0/1 matrix cannot play a causal role in subserving that capacity, but I am insisting that it is not tantamount to that capacity, far from it. In the absence of a robot actually performing in the world, it is just another ungrounded projection onto a bunch of meaningless symbols.

To make the point even clearer: If you took a TTT-scale robot, with all its symbols grounded, and you threw away all the parts of it that interact with the world, reducing it just to the symbols plus, say, the last presymbolic component of the sensorimotor system -- let’s even suppose it’s 0/1 matrices -- you would no longer have a grounded symbol system, any more than you would still have a furnace or a plane if you removed each of their respective "transducers" and "effectors." This is not at all an arbitrary constraint or an idle analogy: What makes a plane a plane is the fact that it can really fly, in the real air. (Not that it DOES fly, but that it CAN fly, that it has the requisite causal powers.) Ditto for a furnace and heating. Well the same is true for a robot and meaning (or thinking, or, to be more specific still, grounding).

Philosophers (and computer scientists) love to think here of the "brain in the vat," cut off from the world, yet still thinking. That's fine, but as I've pointed out many times before, the real brain consists mostly of projections and re-projections of the sensory surfaces, and once you've unravelled and discarded those, you reach the motor projection regions. Discard those and you have very little brain left, and certainly too little to warrant believing that there's still any mind (or meaning) in there (not to mention that the TTT capacity is gone too). What makes some people wrongly believe that what you would have left there is a symbol system? The irresistible homuncular projection that's involved in interpreting symbol systems. But once we realize that that projection is ungrounded then we begin to appreciate what causal role the real sensorimotor part might have been playing in grounding it. And that no amount of interpretation (e.g., of 0/1 matrices) can substitute for that grounding.

To put it as succinctly as possible: Grounding is not just a matter of hooking a symbol system onto transducers and effectors. The transducers and effectors are an essential part of the grounded system.

I will close by answering your specific questions:

> Do I understand correctly: you feel that Jim Geller's program, which
> inputs "pictures" (built of zeros and 1's, using an editor) and simple
> word-strings is NOT properly grounding the learned words to
> sensory-motor events?

Correct. It is merely associating meaningless symbols that are interpretable as words with meaningless symbols that are interpretable as sensory-motor events. Real grounding requires the causal power to interact (TTT-indistinguishably) with real sensory-motor events.

> what WOULD BE adequate grounding:
>
> If the picture had been input by a TV camera?

Not unless the system, receiving input (its proximal projection) from the TV camera aimed at the (distal) world, could behave indistinguishably from the way people do in that world.

> Or drawn on a tablet by a person?

No. And this possibility seems irrelevant.

> Or the system had a robot cart that wheeled the TV camera toward the table?

Since the object is to scale up to TTT-scale robotic capacity, the system would have to have the requisite sensorimotor means for interacting with the real objects, events and states of affairs in the world.

> and aimed the camera? and focussed it and pushed its on-button?

The system's powers better be in the same causal loop. Otherwise it just reduces to interactions between symbolic descriptions of objects and events, which is just the ungrounded "pen-pal" version of the Turing Test (the mere TT) all over again. The TT was what the robotic upgrade (the TTT) was meant to remedy. The pen-pal version -- symbols in and symbols out -- is logically incapable of breaking out of the symbol/symbol circle. Grounding must be direct; not mediated by interpretations or actions that are outside the system. And note that it is the WHOLE system that is grounded, not some subcomponent of it.

> What if the very simple pictures that have already been used were
> actually the result of input from a TV camera? or a robot cart, etc.?

As I said, it is conceivable that a 0/1 matrix could be an actual subcomponent of a grounded system. So what? It is the system that is grounded, not its parts. (By way of a reductio ad absurdum, suppose I am right know thinking -- groundedly, of course -- of a cat. If you took just the part of my brain that betokens that symbol, plus, say, it's immediate connections, e.g., a bit-map, that isolated part would NOT be grounded. Grounding is a systematic property and consists of my capacity to do the many other things I can do with both symbols and the objects in the world that they stand for.)

> I'm keeping things simple to stimulate you to object, and MAKE THEM
> ADEQUATE. If none of these is sufficient, what do humans or other
> animals do in addition? Would a human being who manipulated carts and
> robot arms on the moon or in the next room, using TV cameras, etc. no
> longer be grounded?

This question is not relevant to grounding. We know people are grounded. So extending their senses artificially does not say anything about anything. (Of course extended people are still grounded.) What we want to know is what it is about people that makes them grounded, and one answer is: Whatever it takes to give a robot capacities that are TTT-indistinguishable from those of people. So it's the robot's capacities that are it issue, not an augmented person's.

(It is an empirical question whether reduced human capacities can help in understanding grounding. I am repeatedly reminded that blind, deaf, paraplegic, aphasic and retarded people still have grounded symbols, even though they cannot pass the TTT! True, but we don't know what internal structures and processes are spared in such people, despite their peripheral limitations. So it seems safer to see what it takes to pass the full normal TTT before weakening it in a direction that could reduce it to an arbitrary toy task.)

> If this isn't a good example, do you have any ideas for a better one?
> I'm pressing the concrete since theoretical discussions, although very
> interesting, are very hard to pin down.

The reason a data structure that is arbitrarily dubbed a "sensory input" is not a good starting point is that it is not at all clear that such a data structure has the requisite properties of a real proximal sensory projection, either in the latter's causal interactions with the world of distal objects or in the constraints those unspecified properties would bring to bear on the rest of the system. For it is those unspecified properties -- over and above whatever simple isomorphism inclines us to dub the data structure a "sensory input" in the first place -- that are the real mediators of the grounding. Without them it's still just the symbol/symbol circle mediated by our interprations.

> It seems to me that the sensory/perceptual interface to the "real world
> out there" certainly does introduce an overwhelming richness and
> complexity of information - posing great problems in terms of
> processing and understanding it, and also great potential rewards. I
> like the idea of grounding as needing this great complexity of the real
> world, and also perceptual processes far more complex than
> transduction. But then it becomes a matter of degree-of-complexity and
> there is no sharp line. We could model pieces of the real world inside
> the computer (just as we are already doing when we model a vision or
> language-learning system). Are you arguing for something more?

The difference between real grounding and simulated grounding is not just a difference in degree of complexity. Of course, if we could somehow second-guess what the relevant properties of both the sensory projection and all environmental contigencies were, we could in principle design and eventually pass the TTT with a simulated robot and a simulated world, one that would immediately and transparently generalize to a real robot in the real world (if we ever bothered to build one). By the same token, planes could in principle have been designed and tested exclusively by computer simulating all their relevant properties and those of their environment, so that, as soon as all this symbolic testing was done, the very first prototype plane built according to the information derived from the simulations actually flew successfully.

Yet, highly unlikely though this scenario is, and as far as it might be from an optimal research strategy, it would still be true that, just as the complete simulated plane was not really airborne, only the real one was, so the complete simulated (TTT) robot would not really be grounded; only the real one would be. And the difference would not just be one of degree of complexity (indeed, ex hypothesi, the complexity of the simulation and the real thing would be equal).

Stevan Harnad

PS For those who are interested, the following compressed files are retrievable by anonymous ftp from princeton.edu in directory /pub/harnad

The Symbol Grounding Problem. Physica D 42: 335-346 (1990) [symbol.Z]

Minds, machines and Searle. Journal of Experimental and Theoretical Artificial Intelligence 1: 5-25 (1990) [searle.Z]

Category induction and representation. Cambridge University Press (1987) [categorization.Z]

Other bodies, other minds: A machine incarnation of an old philosophical problem Minds and Machines 1 (1991) [otherminds.Z]

Computational Hermeneutics. Social Epistemology 4: 167-172. (1990) [dietrich.crit.Z]

Lost in the hermeneutic hall of mirrors. Journal of Experimental and Theoretical Artificial Intelligence 2: 321 - 327. (1990) [dyer.crit.Z]

Tim, sorry for my long delay in replying! Just the unrelenting press of obligations.

> From: Tim Smithers
> Date: Thu, 4 Oct 90 21:01:22 BST
>
> Dear Stevan,
>
> I enjoyed your talk here yesterday, and it was good to meet you in
> person--it's now much harder to think of you as being a very good pen
> pal program which interacts with the net :-).
>
> I wonder if I might humbly offer some thoughts on your talk. From
> reading your papers and net exchanges and from now hearing your talk
> I've been wondering if it would be possible to present the symbol
> grounding problem and your proposed possible solution to it without any
> reference to Searle's Chinese Room argument. This might be more an
> academic exercise than a useful thing to do, but it seems to me that
> all the inevitable discussions that go on around Searle's CRA can get
> in the way of presenting the symbol grounding problem and in any
> discussion of possible solutions to it. It also seems to me that
> Searle's CRA is in fact not essential for your argument.

His argument is not essential, but I think it's correct (once it's made clear that it works only against a pure symbol system alleged to have a mind) and I do think it's a particularly revealing instance of the symbol grounding problem. However, I do agree that it tends to cannibalize discussion, so I have in fact taken to compressing or even omitting it in my talks, along the lines you suggest.

> During your talk, when you were presenting what a symbol system
> consists in, you mentioned that some people had suggested that your
> points 2 and 3 might be too strong. Do you now have a modified version
> of this symbol system specification. I ask because I've been using your
> description in some of my teaching, and would like to keep things
> up-to-date.

I acknowledged Paul Kube's point in the published version (available by anonymous ftp from princeton.edu in directory /pub/harnad as compressed file symbol.Z) but am not altogether convinced. If you want to reduce the list to the most important properties, they are (1) arbitrariness of symbol token shape, (2) manipulation purely on the basis of shape, (3) compositionality and (4) systematic semantic interpretability.

> I was surprised about your news that Rosenschein seemed to be denying the
> need to build real robots. I checked my recent net receipts but couldn't
> find this. Could you re-send this exchange.

In Edinburgh I mis-remembered WHICH roboticist had posted that comment to this group. It turns out it was Hans Moravec, not Stan Rosenschein. Let me know if you don't have that exchange and I'll post it to you (or you can retrieve it from the Symbol Grounidng archive in princeton.edu).

> Finally, I wonder what you would say about the following rather rough
> thoughts on grounded symbol using devices?

This is where the SG posting begins:

GROUNDEDNESS, INTRINSIC MEANING AND THE TTT

> If I build a Turing machine, i.e. if I build a real device which has a tape
> (big, but finite), a tape read/write head and associated mechanism for
> moving the read/write head along the tape, I would have a symbol system
> which uses grounded symbols. In this case the necessary transduction is done
> by the read/write and move mechanisms of the machine, and the necessary
> categorisation and identification and subsequent intepretation of input
> signals is built into the device by me--which happens to be quite easy to do
> in this particular example. If you agree this far, would you also agree
> that the same argument can be made for a microprocessor-based computer--a
> programmable Turing machine? In other words, would you agree that for a
> device to be a computer (in the Turing sense), i.e., have systematically
> interpretable states, it must also be a grounded symbol using system? It
> seems to me that this is true and that the meanings of the symbols are
> intrinsic to the machine and derived from the operations it can perform on
> the symbols. The reason for setting up all this is because it seems to me that
> the mistake that many a symbolic functionalist makes is to say that because
> the symbols a computer uses to be a computer are grounded then any symbols
> it is programmed to process must also be grounded. The interesting reality
> is that using a programmable computer, a grounded symbol using device, to
> realise a symbol system (by getting it to execute the right kind of program)
> does not of its self result in another grounded symbol using system. The
> reason is of course that the domain of the meanings of the symbols of the
> new symbol system are not derived (in general) from the operations that the
> machine can perform, but are about things external to the machine and its
> operations on its symbols, and so have to be projected on to those symbols
> by some external agency--us, for example.
>
> All this is another way of saying (I think) that for the symbols of a symbol
> system to have intrinsic meaning (be grounded or derived by composition from
> grounded symbols) that symbol system must be embedded in, and interact with
> in a causal way, the domain which contains the things that its basic meaning
> symbols are about and refer to. In the case of this domain being the real
> world this 'semantic coupling' has to be achieved through a suitable
> combination of transduction, categorisation, identification, and
> interpretation processes (to put it simply). The reason that this cannot be
> achieved by just attaching some sensors and actuators to a symbol processing
> computer is because the semantics of the symbol composing that the symbol
> system engages in is not independent of the grounding processes required,

> and so not just any symbol system can necessarily be grounded. To get it
> right we have to build bottom up, as you say. That's the way the
> constraints operate in this world.

I think I agree with a lot of this, except, as a terminological point, I would not use "grounded" to refer to an implemented Turing Machine. If "grounded" merely meant "causally embedded in the world," then every phsyical object would be grounded, including an implemented symbol system (and only unimplemented, abstract formal symbol systems would be "ungrounded" -- though even they would be systematically interpretable).

Grounding becomes a substantive matter when you can speak about the interpretations of the symbols in a symbol system: Are THEY grounded? And it seems clear that this question only arises when the implemented Turing Machine is interpretable as doing something other than what it is LITERALLY doing (register 3 active now); for in that trivial sense a chair too is grounded (sitting right here now). I also think it's a bit of a cheat on systematicity if the semantics in question is just a bunch of identity statements.

But maybe you mean something else, something I WOULD be prepared to disagree with. Maybe you mean that a Fortran payroll program or statistics program is grounded; that it's really "about" my salary and the gross national product, or even the quantity 500,000. I would deny all of those. The program is interpretable as being about payrolls or numbers, but that interpretation is ungrounded: It is mediated by our minds. Now I realize that a Fortran program has no ambition to have a mind of its own, but that's just what I would take the claim that ordinary computers are grounded to be implying. Even if the computer were so causally embedded as to actually be controlling the physical disbursement of my money to me I would say that its symbols were not about money (or about anything, for that matter).

For me the problem of grounding is this: If the symbols in a system independently ("intrinsically") mean what they mean (rather than meaning them only in virtue of the fact that we so interpret them) then there must be a coherent and systematic link between what the system "says" (or can be interpreted to say) about the world and what it does in the world. Now this is where the inadequacy of toy tasks and the necessity for the TTT comes in: The symbols in an IBM payroll program, even a "dedicated" one that is directly wired to all employees' time-clocks as well as to their bank accounts, are still not ABOUT money in the sense that my conversation is about money (when it happens to be). Why? Well although you can point to the causal link between what the computer can be interpreted as symbolizing that it will pay me and what it actually pays me, there are countless other systematic relationships with the meaning of money and work that are absent from this "toy" system yet are clearly part of the meaning of money and work.

Is the proper conclusion then that such a "dedicated" toy symbol system is indeed grounded, only very primitive and very stupid, meaning something by its symbols, but something very primitive and simple (like the housekeeper with Alzheimer's disease described in Steve Stich's book, whose only surviving proposition was "They shot McKinley")? I think not. I think only TTT scale systems mean anything at all. Sub-TTT systems, even dedicated ones, hard-wired (or even servo-linked, with learning capacity) to the referents of their symbols, are as ungrounded as pure symbol systems.

My reason has to do with consciousness and degrees of freedom. I think a system must have a mind in order to be grounded. Having "intrinsic meaning" means HAVING MEANING TO THE SYSTEM ITSELF, and for that there has to be somebody home in there. And I don't believe "somebody being home in there" is a property that admits of degrees (as it clearly would have to if every dedicated system were grounded and it was just a question of how "smart" the grounding was).

Now the TTT could, logically speaking (and in reality), be either too weak (falsely implying that mindless TTT-scale systems are mindful) or too strong (falsely ruling out as mindless sub-TTT systems that are mindful). But I’m betting it’s just right. Not by definition (that would be absurd), nor even by hypothesis (because, as I said, whereas the TT is vulnerable to Searle’s "periscope," I can’t imagine an equivalent strategy for falsifying the TTT), but for methodological reasons: It’s the only way to bet, and the blind watchmaker (evolution) is as blind to TTT-indistinguishable differences as we are. Yet organisms seem to have evolved into their robotic worlds with minds. So it must be something about the performance demands of their worlds (in contrast to that of the dedicated payroll computer) that ensured (if so it did) that they would be mindful rather than mindless. The TTT exacts precisely the same performance demands that evolution did.

But the TTT could conceivably be a false guide. What would follow then? Well, then clearly a grounded system would not be equivalent to a mindful system. The systematic symbolic/sensorimotor coherence required by the TTT might pick out SOME important property, but it would not have anything to do with whether symbols mean something TO the system, because there would be no one home in the system for the symbols to mean anything to. In that case, again, apart from whatever performance dividends might accrue from "grounding" a system in this mindless sense, it’s clear that there would no longer be much difference between a mindless ungrounded system, in which all meanings are just projected by the external mind of the interpreter, and a mindless grounded system, which would meet the performance criteria for interacting coherently with the objects and states of affairs that its symbols allegedly mean, yet, with no one home in there to mean them, it would still just be a matter of interpretation, only now the external mind can project an interpretation not only on the system’s symbols, but also on its actions, and the relations of its symbols and actions to the world!

A logical possibility. One that could drive a wedge between "groundedness" and "intrinsic meaning." But not one I’m betting on. Not every dedicated computer or robot is grounded: Only the ones with TTT-scale performance capacities can narrow the functional alternatives to those faced by evolution, in which the outcome converged on was evidently: somebody home in there for the world to mean something to.

Stevan Harnad

To: Symbol Grounding Discussion Group

IS THERE A LEVEL BETWEEN THE TURING TEST (TT) AND THE TOTAL TURING TEST (TTT)?

> Date: Thu, 20 Sep 90 18:30:56 PDT
> From: nrouquet%gringo.usc.edu@usc.edu (Nicolas Rouquette)
> Subject: Re: Simulated robots/Simulated worlds
>
> Date: Thu, 20 Sep 90 17:33:50 EDT

> From: Stevan Harnad
>
> Here's an experimental prediction: Only a system capable of passing the
> TTT could pass the TT. It would follow from this prediction that since
> a pure symbol cruncher could not pass the TTT (because of the empirical
> difference between computers and transducer/effector/analog systems), a
> pure symbol cruncher could not pass the TT.
>
> This argument assumes that there is nothing between TTT capability and
> TT capability. How strong is the knowledge available to support such a
> claim? This seems to me an important question: Should there be the
> possibility of a class of machines between TTT and TT capability, then
> lots of the past discussions would have invalid claims as passing the
> TT would not necessarily require full blown TTT capabilities.
>
> Nicolas Rouquette Computer Science USC

The TT (Turing Test) calls for performance indistinguishablity in all of our symbolic (verbal) capacities. The TTT (Total Turing Test) calls for performance indistinguishability in all of our robotic (symbolic AND sensorimotor) capacities. What is in between? Some of those capacities instead of all? But some is tantamount to distinguishability then: We're not inclined to say: He's just like a person in how he plays chess and orders food in a restaurant but not in any other respect. On the contrary, subtotal capacities are precisely the clue to FAILURE on the Turing Test because of distinguishability from ourselves.

There do exist two principled ways to restrict the TTT, but neither seems methdologically viable: You could ask for a nonhuman version (an ape TTT or a rodent TTT), but unfortunately we are neither ecologically nor empathically capable of judging whether or not the candidate is indistinguishable from the real organism in any other species than our own. Or you could ask for a clinical human version (TTT indistinguishability from a person who is blind, deaf, aphasic, apraxic, paraplegic, retarded, etc.). Here the problem is that the TTT is a performance test, and we do not know what performance capacities are preserved in such patients, but masked by their peripheral damage. So this too would be to weaken arbitrarily an already fallible performance criterion, and to increase the degrees of freedom of cognitive theory when they already exceed the normal underdetermination of scientific theories by their data. (But if the underdetermined clinical TTT risks leaving out a performance capacity that is relevant and essential, just as a toy task does, then the overdetermined TTTT (calling for complete neuromolecular indistinguishability) risks sidetracking us toward irrelevant and inessential properties.)

Stevan Harnad

SIMULATED SENSE DATA VS. SIMULATED SENSORS

> Date: Mon, 14 Jan 91 00:46:03 PST
> From: Dr Michael G Dyer
> Subject: grounding
>

> It appears that the term "grounding" can be used in two different
> ways:
>
> (1) that words, concepts, etc. be "grounded" in perception/motion,
> i.e. that there not just be "symbols" only, but that images, and
> trajectories and other kinds of computations, related to vision,
> coordination, etc. be included.
>
> (2) that there exist PHYSICAL sensors (transducers)
>
> I’m happy with use #1 above, which does not seem controversial at all,
> and various forms of connectionist research is directed at merging
> symbolic-type representations with sensory-type representations.
>
> I’m not happy with use #2 of the term "grounding", however. This use
> seems theoretically unnecessary and is based on an incorrect acceptance
> of Searle’s arguments. "Grounding" should always be with respect to
> some environment. An environment need not be physical -- it can be
> simulated. A simulated robot could then be "grounded" with respect to
> that simulated environment. How well its "intentionality" matched our
> own will then depend on both the robot’s complexity AND the complexity
> of its environment. Theoretically, at least, if both the robot and the
> simulated environment approach that of our own, then we should grant
> intentionality to the robot, WITHOUT the need for PHYSICAL sensors.
>
> Anyone who wants to read more about this point of view may request
> the following reprint from dyer@cs.ucla.edu
>
> Dyer, M. G. Intentionality and computationalism: minds, machines,
> Searle and Harnad. Journal of Experimental and Theoretical Artificial
> Intelligence. Vol. 2, 303-319, 1990.

Actually, based on the latest postings, there are not two, but THREE ways one could use the word "grounding." Each has some interesting consequences.

(1a) Purely Symbolic "Grounding": Let us call those symbols "grounded" that (a) can be systematically interpreted as referring to objects, actions, states and events in the world and that (b) co-occur in a program that also contains other symbols that can be systematically interpreted as being those objects, actions, states and events in the world.

[This sense of "groundedness" would presumably contrast with symbols that were "ungrounded" because (b) was missing. But if one understands the symbol grounding problem -- which, briefly put, is that symbols alone, be they ever so systematically interpretable, are NOT grounded: grounding requires the right causal connection with the actual referents of the symbols, not just more symbols that can be interpreted as the referents of the symbols -- then one sees that if anything deserves the name of "pseudogrounding" then (1a) does. To see this, note that it is the INTERPRETATION that we are trying to ground, and an interpretation cannot be grounded in yet another interpretation, for that is just the hermeneutic circle.]

(1b) Robotically Implemented Symbolic Grounding: An alternative that is similar to (1a) but differs in one essential respect is the one discussed in the last few postings: Supposing we manage to capture (symbolically, and UNGROUNDEDLY) all the relevant properties of both the robot we wish to build, and its environment. Suppose the proof of the fact that we had indeed symbolized them all was in the outcome that, using nothing but what we have learned from the simulation, we can immediately build a real robot that does successfully in the real world what the simulated one did in the symbolic world. What would be true here would be (i) that the symbols in the REAL robot would be grounded and (ii) that its simulation had included the full blueprint for grounding. It would NOT be true that the simulation was grounded, any more than it would be true that a simulated airplane in a simulated environment was airborne. Grounding depends completely on the real transducers and effectors (and any other analog processes in between) and their causal properties in the real world; the nonsymbolic components are essential parts of the grounded hybrid nonsymbolic/symbolic system.

(2) Nonmentalistic Robotic Grounding: A second alternative would be to ignore the mental aspects of the distinction between (i) symbols with "derived meaning" (symbols that are interpretable as meaning X because I, a being with a mind, can and do systematically interpret them as meaning X) and (ii) symbols with "intrinsic meaning" (symbols that mean X in and of and to themselves, because the symbol system has a mind of its own) and simply use "grounded" to refer to dedicated symbol systems (e.g., robots) whose symbolic and sensorimotor behavior in the world can be given a coherent joint interpretation: Robots, be they ever so simple, that can physically pick out and interact with the real objects, events and states of affairs in the world that their symbols can be systematically interpreted as referring to.

[Note that (1b) and (2) are equivalent, but (1b) is assumed to have been arrived at via simulation alone, and to have a simulated counterpart that captures symbolically all the relevant functional properties of the robot and the world.]

(3) Mentalistic Robotic Grounding: The third alternative is that to have grounded meanings, unmediated by external interpretations, a symbol system must have a mind to which its symbols mean what they mean. This would call for a robot whose performance capacities in the real world were as powerful as our own (i.e., a robot capable of passing the Total Turing Test [TTT]).

I favor (3), but note that if in reality the TTT is not sufficient to guarantee that the robot has a mind, then (3) becomes equivalent to (2). [Tim Smithers holds to (2).]

So here is where my difference with Mike Dyer resides: He believes that even (1a) is "grounded," whereas I believe that if the problem of symbol grounding has any substance at all, it is that it rules out a purely symbolic solution like (1a). It’s precisely the symbol/symbol circle, held together only by its amenability to a systematic interpretation projected by external minds, that symbol grounding proposals are attempting to break out of! That’s also the "Hermeneutic Hall of Mirrors," which I have argued Mike Dyer is lost in.

Stevan Harnad

> Date: Mon, 14 Jan 91 01:00:25 PST
> From: Dr Michael G Dyer
>
> you state that, EVEN IF ALL aspects of our physics were captured, the
> simulated plane would not really be flying.
>
> How do you know whether or not our own universe is not simply itself a
> simulation? we could be inside some 10**150 processor "hypercube" and
> all "motion" is simply the replication of patterns displaced in space
> (a la Game of Life). YOU CAN'T TELL, can you? Suppose we are. We still
> retain our own "intentionality". So it doesn't appear that "simulating"
> something immediately destroys its capacity for motion or
> intentionality. At least one can't make the argument by beating on the
> term "simulation". WE don't know if we're a simulation. If we are, then
> simulation is irrelevant to intentionality. If our universe is not a
> simulation, then it's still up in the air as to whether or not some
> situated intelligence within some simulation we create has
> intentionality itself.
>
> Anyone who reads Harnad's "Lost in the hermeneutic hall of mirrors"
> should ALSO read:
>
> Dyer, M. G. Finding lost minds. J. Expt. Theor. Artif. Intel. Vol. 2,
> 329-339, 1990.
>
> reprints may be requested of dyer@cs.ucla.edu

In "IF ALL aspects of our physics were captured," "captured" is equivocal: I mean, of course, captured SYMBOLICALLY in a form that is Turing Equivalent to and systematically interpretable as the physical properties in question.

There are two senses of simulation that invariably get conflated when one gets lost in the hermeneutic hall of mirrors created by projecting our interpretations onto symbol systems and then simply reading them back again, by way of confirmation:

(A) It is logically possible that I am a real physical brain, but that all the inputs to my (real) senses are created by a sensory simulator: It is not the shadow of a real apple that falls on my retina, but a pattern created by a CRT; I am not really spinning in circles, it is just an oscillator hooked directly to my semicircular canals. This is the sense of "simulation" that is meant when we refer to a flight simulator or video game that simulates driving for children. Note that what you have there is a real (grounded, mindful) human being placed in a world in which the inputs to his senses are simulated ones.

(B) The second sense of simulation is the one in (1a) above, in which it is not only the inputs to the sense organs that are simulated, but the robot itself. Now please make an important observation: This is not an implemented robot (1b) in a simulated environment in the sense of (A). It is a pure symbol system (a computer, say) in which some of its symbols are interpretable as being the environment and some of its symbols are interpretable as being the robot. This is exactly the same as the example of the simulated aiplane discussed in prior postings: both the airplane and the air are symbolically simulated. It is not a real person or airplane in a simulated flight environment, or vice versa. It’s symbols (and interpretations) only.

Now it’s a universe that consists of (B) alone that Mike Dyer wants to ask us to believe the REAL one might be! A huge computer running a program whose states are systematically interpretable as being you and me. In my view, the inclination to take such a sci fi scenario as a serious possibility is a clear symptom of being lost in the hermeneutic hall of mirrors that we create when we uncritically project interpretations onto symbol systems and, impressed by their ability to sustain the interpretations coherently and systematically, we forget the distinction between symbol and reality, and that the real source of all the meaning is our own minds and not the symbol system (which, denuded of interpretation, is really just "squiggle-squiggle, squoggle-squoggle").

What is being conflated is (i) the capacity of sensory simulations to fool real, mindful me or a TTT-scale robot, on the one hand (these might be generated by a symbol system driving a video screen, which is then viewed by me or the TTT-robot, or by a symbol system driving devices that directly stimulate my nerve endings or the robot’s transducers, or even deeper loci in my brain or the equivalent parts of the robot’s innards) and, on the other hand, (ii) a pure symbol system, part of which is interpretable as sensory input (perhaps it could literally be the symbol system that drives the video in (i)) and part of which is interpretable as me or the robot (perhaps even the one in (1a) above that, when actually implemented as a robot with transducers and effectors, would be able to pass the TTT).

It should be quite obvious where the logical error (or the arbitrary leap into the merely logically possible but vastly improbable) occurred: Unless we PRESUPPOSE that the only thing that matters is the symbol system -- that the only thing that makes me and the TTT-scale robot have a mind is that we are the implementations of the right symbol system -- then there is a crucial difference between (i) and (ii), namely, (ii) is just a symbol system plus projected interpretations whereas (i) includes nonsymbolic properties such as transduction. And the symbol grounding problem suggests that this presupposition is simply wrong, because the interpretation of the symbol system is ungrounded -- it’s hanging by a skyhook.

Note also that if we WERE all just states in a big symbol system rather than the physical organisms with real transducers and effectors in the real world of objects that we imagine we are, then all of physics (including the physics of flight, heat, motion, matter, energy) would all be pseudophysics, because there would be no real flight, etc., just its symbolic simulation in this vast symbol system in the sky. But if that were so, how could you even help yourself to the (real?) physical properties you need to implement the cosmic computer that must be imagined in order to spin such a sci fi fantasy in the first place? In other words, I even find the fantasy itself incoherent, arbitrarily picking and choosing as it does among the physical properties that it allows to be real vs. merely simulated.

I prefer not to enter this hall of mirrors, retaining instead the perfectly valid distinction between simulation and reality, and between pure symbol systems that are systematically interpretable as mindful, but aren’t, and real robots grounded in the real world, that are.

Stevan Harnad