University of Birmingham Graduate School Seminar, Tuesday 24th June 2008

Series in Computational Approaches to Biological Research

Evolution, development and modelling of architectures for intelligent organisms and robots.

Aaron Sloman

http://www.cs.bham.ac.uk/~axs/

(Based partly on work with Jackie Chappell)

These slides will go into my 'talks' directory:

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#biosem

Abstract

Evolution, development and modelling of architectures for intelligent organisms and robots

Depending on time available, and how the discussions go, I'll try to say something about:

- How philosophy and biology are mutually relevant;
- The importance of what Dennett describes as the design stance as opposed to other stances;
- Why so much in biology is information-processing;
- How artificial intelligence extends our ideas about information-processing;
- Some features of the design-stance required to understand information processing systems, including the role of virtual machines;
- Recent work I've done with Jackie Chappell on nature/nurture tradeoffs;
- Some ideas about information processing architectures in human-like animals and machines living in a changing world of 3-D structures and processes;
- Requirements for modelling tools, partly illustrated using our SimAgent toolkit.

Bioinformatics: Three Kinds

If we interpret “Informatics” in the new broad sense covering all aspects of information-processing, including theory and applications in both natural and artificial information-processing systems, then “bioinformatics” acquires a new broad connotation covering at least the following:

- Development of techniques and mechanisms for acquiring, storing, handling and applying data from observation or study of biological systems. (Scanners included?)

  This is how many people still think of “bioinformatics”.

- The study of information-processing systems that occur in the biosphere, including information-processing
  - in individual organisms,
  - in groups of organisms

    (e.g. hunting packs, grazing herds, swarms, flocks, hives, nests, termite cathedral construction),
  - in ecosystems
  - in evolution.

- The construction of working models of such information-processing.

This talk is mostly about the last two, though many people think of bioinformatics as concerned with the first category.

My interest is in information-processing architectures in intelligent individual organisms

e.g. nest-building birds, hunting and tree-climbing mammals, primates, including humans, and in the problems of designing robots with simiar capabilities.

Different forms of information-processing

The systems studied, the models used, and the information-processing techniques deployed may have different features and functions:

- Manipulation of atomic numerical data, or possibly a mixture of numerical data and some symbolic atoms (e.g. names of species, continents, seasons, etc.)
- Interpretation and manipulation of structured information in many forms

  E.g. sentences, plans, trees, networks, images, proofs

  The structures may vary in complexity, may combine different sorts of information (numerical and non-numerical) and some of them may be changing structures, e.g. the information structures in a system that learns.
- Control of physical systems, e.g.:
  - animal bodies,
  - things in the animal's environment,
  - experimental apparatus in biological experiments
  - robotic models
- Sensing and interpretation of ongoing processes involving various combinations of the above – e.g. modelling processes of perceiving or controlling an action, e.g.
  - Walking, running, jumping
  - Biting, picking up, dissassembling, assembling
  - Gesturing, producing a sentence
  - Dancing, fighting, hunting, collaboratively building a house.
- Control of internal processes, e.g. thinking, planning, imagining, deciding...

Intelligent design by evolution and by biologists

Many kinds of information-processing produced by biological evolution (aided by learning and development) are far more sophisticated than anything we currently know how to produce in artificial systems.

How can we learn about and understand such systems?

- It is tempting to think that more observation of natural systems will reveal what they do and how they do it.
- However, unless biological researchers learn how to design, implement, test, and debug complex working systems themselves, they are unlikely to know what to look for or to understand what the observed organisms are doing or why, and even less likely to find out what mechanisms are used.
- So it is important for more people studying animal behaviour, human intelligence, etc. to gain experience of designing and testing working animal-like or human-like systems (starting simple of course.)
- Unfortunately many of the most widely used programming languages and programming courses teach only a subset of the skills, techniques and concepts needed, e.g. emphasizing numerical programming techniques.

In short: There is a need for intelligent designers in the biological sciences

(including psychology, neuroscience, social science...).

Philosophy, AI and Biology

- Philosophical analysis can help people in all disciplines get a clearer understanding of conceptual issues (e.g. what do we mean by “information”, “life”, “inheritance”, “cognition”, ...?)
- Doing philosophy can help designers of intelligent systems be clear about the goals they are aiming for, and the criteria by which their work should be evaluated.
- Learning about designs for intelligent information processing systems helps to shed light on some old philosophical problems, e.g.
  - problems about the relationships between mind and body
  - problems about free will and determinism
  - how biological evolution could have produced minds
  - how varied the space of possible minds is
  - what computation is (much more than what computers do)
  - what science is and is not

  And several more.

Some possible things to read

The following provide more information about the overlaps:

- Talk 10: What is Artificial Intelligence?

  http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#whatsai
- Talk 13: Artificial Intelligence and Philosophy

  http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#aiandphil
- Other presentations: http://www.cs.bham.ac.uk/research/projects/cogaff/talks/
- *The Computer Revolution in Philosophy: Philosophy science and models of mind* (1978)

  http://www.cs.bham.ac.uk/research/projects/cogaff/crp

Obvious motivations for studying AI/Robotics

Motivations can be

- **practical** or
- **theoretical** (including science and philosophy)

The most obvious and common motivations for building AI systems are **practical:**

- Solving existing practical problems

  (e.g. improving automated assembly, or automated advice, sales, booking, or entertainment systems)

- Solving anticipated practical problems

  E.g. providing future domestic robots to help elderly and infirm, or future robot guides to public buildings (galleries, hospitals, etc.)

  A robot companion for me when I am older????

  Robots to assemble space stations.

- Providing modelling tools for other disciplines, e.g. neuroscience, psychology, biology, social sciences, education:

  E.g. helping them formulate their theories in a runnable form.

Less obvious Motivations for studying AI/Robotics

Less obvious motivations: **expanding knowledge for its own sake.**

- Deepening our understanding of varieties of information processing systems: natural and artificial.

  This includes formulating new kinds of questions that psychologists, neuroscientists, biologists, philosophers do not usually think of.

  E.g. questions about information processing architectures, forms of representation, mechanisms.

  ESPECIALLY Questions about varieties of virtual machines, what they are useful for, and how they can be implemented – in brains or other kinds of physical machines.

- Making progress with old philosophical problems

  by providing new conceptual tools

  for articulating the questions and previously unthought of answers

  Including tools for demonstrating and testing philosophical theories

  Example:

  I originally got into AI because I wanted to show why Kant's philosophy of mathematics was correct and Hume's wrong – and eventually I realised that that goal required me to learn how to design a working mathematician – starting from a baby mathematician seeing shapes and learning to count! See chapter 8 of The Computer Revolution in Philosophy (1978)
  `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`
  Alas it proved much more difficult than I had anticipated – we still are not close!

How philosophy can contribute: consciousness

Several books and conference reports on “machine consciousness” have already appeared and no doubt many more will
(e.g. AAAI’07 Fall Symposium)

Much recent work by AI researchers on consciousness assumes that “consciousness” is a unitary concept, requiring a unitary mechanism.

Philosophical analysis can show that the ordinary notion that we all understand is a mish-mash of inconsistent concepts of different sorts.

Example:

- you are unconscious when you are asleep
- when you are dreaming you are asleep
- you are conscious when you are frightened
- when dreaming you can be frighted by a hungry lion chasing you

How philosophy can contribute: consciousness

Several books and conference reports on “machine consciousness” have already appeared and no doubt many more will
(e.g. AAAI’07 Fall Symposium)

Much recent work by AI researchers on consciousness assumes that “consciousness” is a unitary concept, requiring a unitary mechanism.

Philosophical analysis can show that the ordinary notion that we all understand is a mish-mash of inconsistent concepts of different sorts.

Example:

- you are unconscious when you are asleep
- when you are dreaming you are asleep
- you are conscious when you are frightened
- when dreaming you can be frighted by a hungry lion chasing you

So, you can be both conscious and unconscious at the same time???

This is just one of many indications that our notion of “consciousness” is muddled.

Owen Holland gives some more here `http://cswww.essex.ac.uk/staff/owen/adventure.ppt`

Philosophical analysis can show

- There is no one thing referred to by the noun ‘consciousness’
- There is no one thing whose functions, evolution, brain mechanisms, (etc.) need to be explained.
- There is a collection of very different mental states and processes that can be described using the adjective ‘conscious’.

In philosophical jargon “consciousness” is a “cluster concept”.

Analysing the cluster of sub-concepts helps to clarify the goals of research in AI.

For more on this see
`http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200302`
A. Sloman and R.L. Chrisley, 2003, Virtual machines and consciousness,
Journal of Consciousness Studies,

Similar comments apply to ‘autonomy’, or ‘free-will’:
another muddled mish-mash concept.

See `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/four-kinds-freewill.html`
Four Concepts of Freewill: Two of them incoherent

Artifical systems can contribute to conceptual analysis

Many concepts needed for describing living things, including their mental processes are very hard to analyse: e.g. ‘learning’, ‘consciousness’, ‘motivation’, ‘emotion’, ‘cognition’.

- If we explore a wide variety of more or less sophisticated information-processing architectures combining many different mechanisms, with diverse functions, we can demonstrate how some of the capabilities they have mirror and explain certain human (and animal) capabilities.
- We can then define new theory-based concepts in terms of states and processes that can arise when such architectures work.
- We replace old, obscure ambiguous concepts with new architecture-based concepts.
- Compare the effect of new discoveries about the atomic structure of matter:

  The periodic table of the elements.
- A deep new theory can revise our ontology.

  AI architectures can generate new “periodic tables” of types of mental processes.
- AI has already begun to revise our ontology for mental states and processes by showing us new, previously unimagined, subdivisions:

  e.g. different sorts of learning, different levels of control; different functions for perception.

A major concept developed recently

One of the most important new concepts to come out of research and development in computing since the mid 1950s is the concept of a **virtual machine** that can be active and do things, while running on another machine, which could be either a physical machine or a lower level virtual machine.

Examples of virtual machines running on millions of computers:

- Operating systems
- file management systems
- backup systems
- networking systems
- web browsers
- email systems
- word processors
- spelling checkers
- security mechanisms
- viruses!

How many virtual machines are known to be implemented in biological organisms, or collections of organisms?

Some virtual machine demos

I shall present live demos of the sort shown in video recordings here:

`http://www.cs.bham.ac.uk/research/projects/poplog/figs/simagent`

Demo 6: shows two toy ‘emotional’ agents moving around, reacting to what they ‘observe’ in the environment, including how close they are to their ‘desired’ targets, whether they have been moved forcibly by the mouse, whether there are obstacles in the way, whether the target has been moved, whether they encounter the other object.

The agents not only produce reactions shown by changes in their speed of movement and the ‘expression’ displayed in a face picture, they are also able to report verbally on their changes, e.g.

> I feel glum because ...
>
> I feel surprised because ...
>
> I feel happy because ...

This really is just a toy teaching demo (with all the code available as part of the SimAgent toolkit) but it illustrates points about virtual machines used later in this talk.

In particular, there are clearly causal interactions between events going on in the virtual machines, and also between the physical environment and events going on in the virtual machines.

A change in the virtual machine (e.g. the “current feeling” becomes surprise) can cause a physical change on the screen. It also causes changes in the physical processes in the computer.

The SimAgent toolkit is described here:
`http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html`

Virtual machines and an old philosophical problem:

What is the relation between mind and body?

- **Mental entities, states and processes seem to be very different from physical entities, states and processes: can we explain the differences and their relationships?**
- When you travel in a train your physical components (e.g. teeth, heart) travel at the same speed, but it seems incorrect talk about your experiences, thoughts, desires, feelings, memories travelling with you: they don’t have locations and therefore cannot move through space.
- If a scientist opens you up, many parts can be inspected and measured, but no thoughts, desires, feelings, memories can be detected and measured using physical devices (though brain processes related to them can be measured).
- Any of your beliefs about your physical environment can be mistaken but certain beliefs about your mental state cannot be mistaken; e.g. believing that you are in pain, that you are having experiences. (Also brain states and processes cannot be mistaken: they merely exist.)
- This leads to puzzles about how such mysterious, ghostly items can be associated with physical bodies.
- Some philosophers have even argued that mental states are all illusory and don’t exist at all.
- If mental processes do exist how can they cause physical events, like human actions, to occur?

That’s a very crude and incomplete summary of a vast amount of philosophical discussion.

We now show how to get some things clearer.

Supervenience and the mind-body relation

Some philosophers have tried to explain the relation between mind and body in terms of a notion of ‘supervenience’:

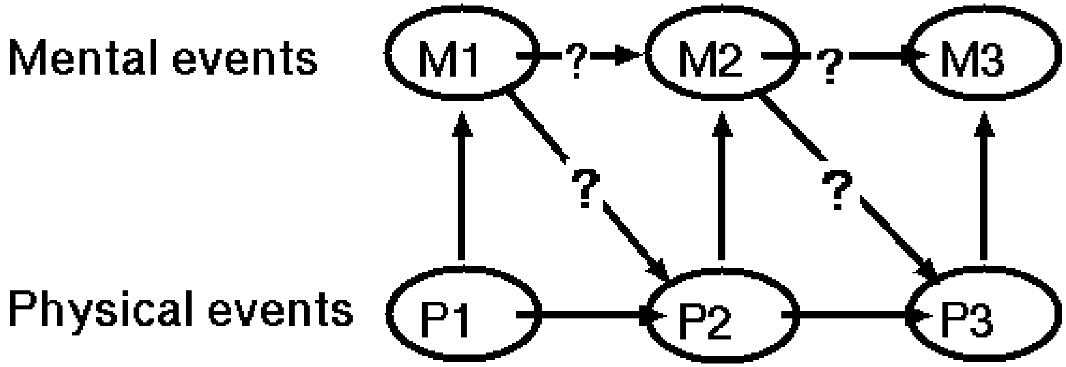
Mental states and processes are said to supervene on physical ones.

(Ideally we should look at several varieties of supervenience, and attempts to produce precise definitions.)

Many problems about that relationship: can mental processes cause physical processes?

How could something happening in a mind produce a change in a physical brain?

In the diagram, think of time going from left to right:



If previous physical states and processes suffice to explain physical states and processes that exist at any time, how can mental ones have any effect?

How could your decision to come here make you come here – don’t physical causes (in your brain and in your environment) suffice to make you come?

N.B. Exactly the same problem arises with virtual machines in computers: how could a running spelling checker (a virtual machine) cause physical changes on the screen?

We (sort of) know how artificial VMs work

Mental functions in organisms are very hard to study: their states and processes and especially their information contents are no more accessible to physical observation and measurement than the states and processes and information contents of virtual machines in computers.

A difference is that the artificial VMs are designed, implemented, tested, debugged, repaired, extended, by human software engineers.

Few people understand all the details: functioning computer VMs depend on very complex interactions between subsystems designed and built independently by:

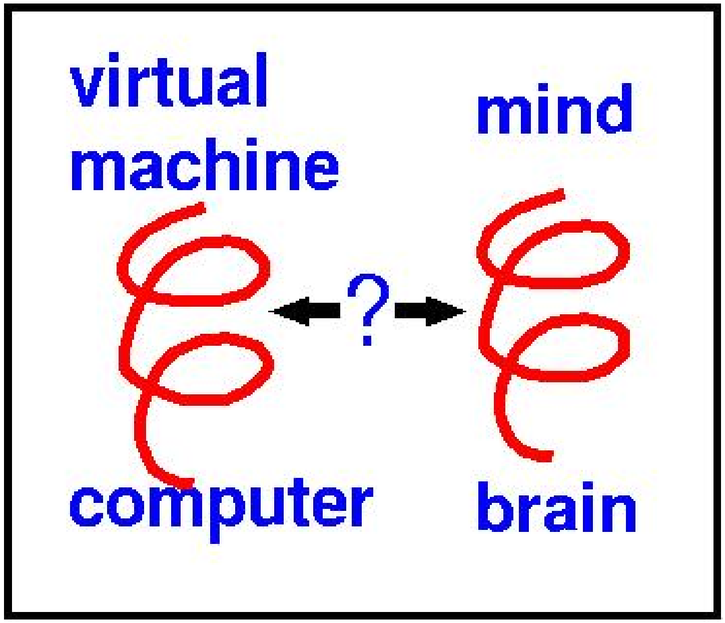
- electronic engineers and digital circuit designers
- firmware/microcode designers
- designers of different components of operating systems
- designers of hardware interfaces and their drivers
- designers of programming languages, and compilers and interpreters for programming languages
- designers of packages using one or more programming languages
- designers of packages using other packages

The upshot is that all these systems (mostly) work together so as to initiate, maintain, protect, and use collections of enduring structures and processes that interact in principled ways, some of them virtual machines interacting with physical components.

Giving a detailed philosophical analysis of the varieties of causation involved is a non-trivial task, however. (Compare autopoesis in biology?)

Biological systems are far more resistant to “hardware” faults and changes, however.

What we have learnt about virtual machines (e.g. programs running on computers), provides new ways of thinking about this – especially AI virtual machines



Many people have explored this analogy, but when philosophers use over-simplified ideas about virtual machines, they produce over-simplified theories and concepts.

Not only philosophers: the general public, and many scientists, are misinformed about this.

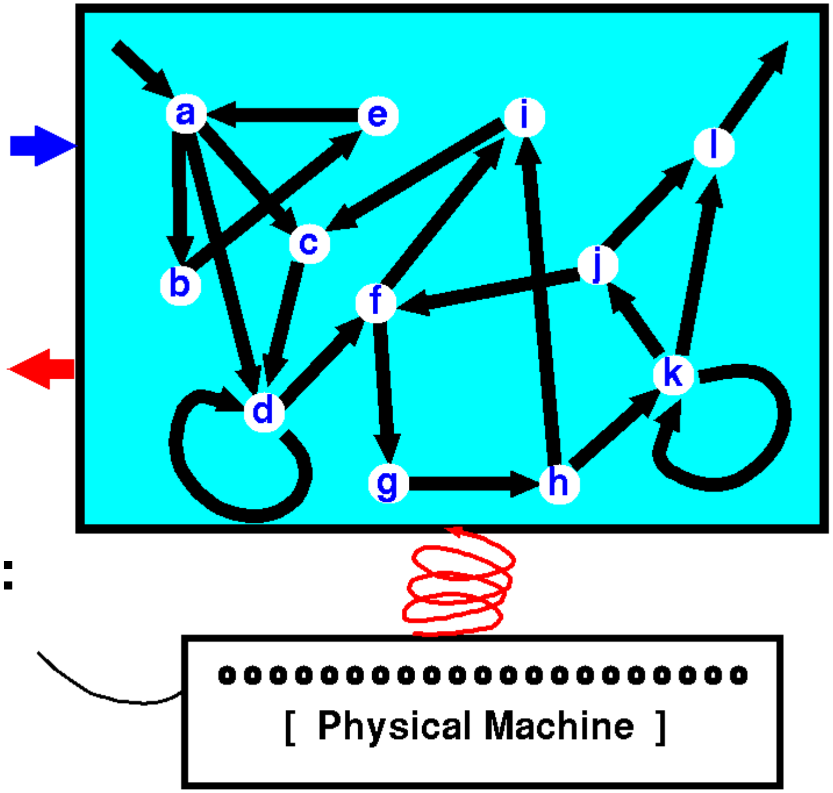
How many people think of virtual machines: Finite State Machines (FSMs)

(e.g. Ned Block once)

Virtual machine:

Implementation relation:

Physical computer:



The virtual machine that runs on the physical machine has a finite set of possible states (a, b, c, etc.) and it can switch between them depending on what inputs it gets, and at each switch it may also produce some output.

This is a fairly powerful model of computation: but it is not general enough.

A richer model: Multiple interacting FSMs

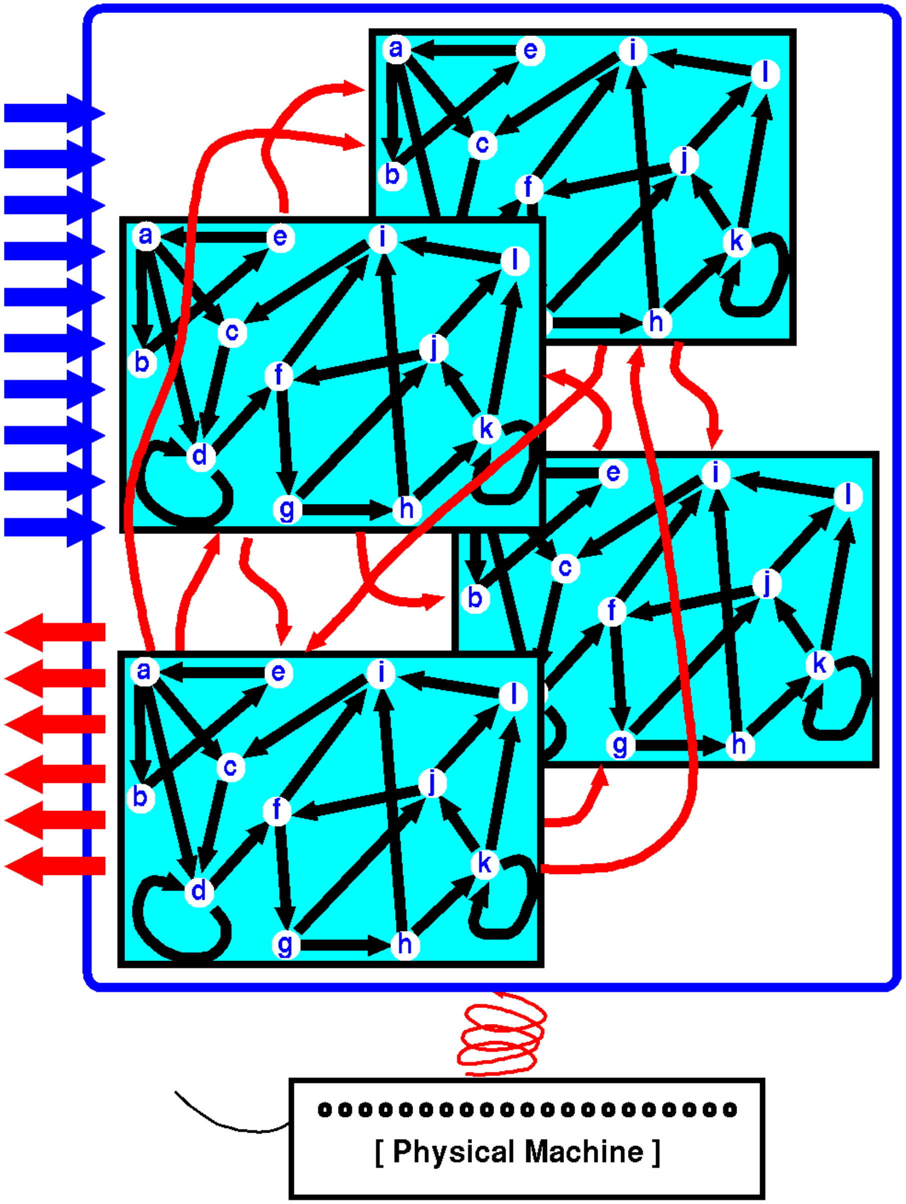
This is a more realistic picture of what goes on in current computers:

There are multiple input and output channels, and multiple interacting finite state machines, only some of which interact directly with the environment.

You will not see the virtual machine components if you open up the computer, only the hardware components.

The existence and properties of the FSMs (e.g. playing chess) cannot be detected by physical measuring devices.

But even that is an oversimplification, as we'll see.



First, a possible objection: Computers are serial

Some will object that when we think multiple processes run in parallel on a single-CPU computer, interacting with one another while they run, we are mistaken because only one process can run on the CPU at a time, so there is always only one process running.

This ignores the important role of memory mechanisms in computers.

The different software processes can have different regions of memory allocated to them, and since those endure in parallel, the processes implemented in them endure in parallel, and effect one another over time. In virtual memory systems, things are more complex.

It is possible to implement an operating system on a multi-cpu machine, so that instead of its processes sharing only one CPU they share two or more.

In the limiting case there could be as many CPUs as processes that are running.

By considering the differences between these different implementations we can see that how many CPUs share the burden of running the processes is a contingent feature of the implementation of the collection of processes and does not alter the fact that there can be multiple processes running in a single-cpu machine.

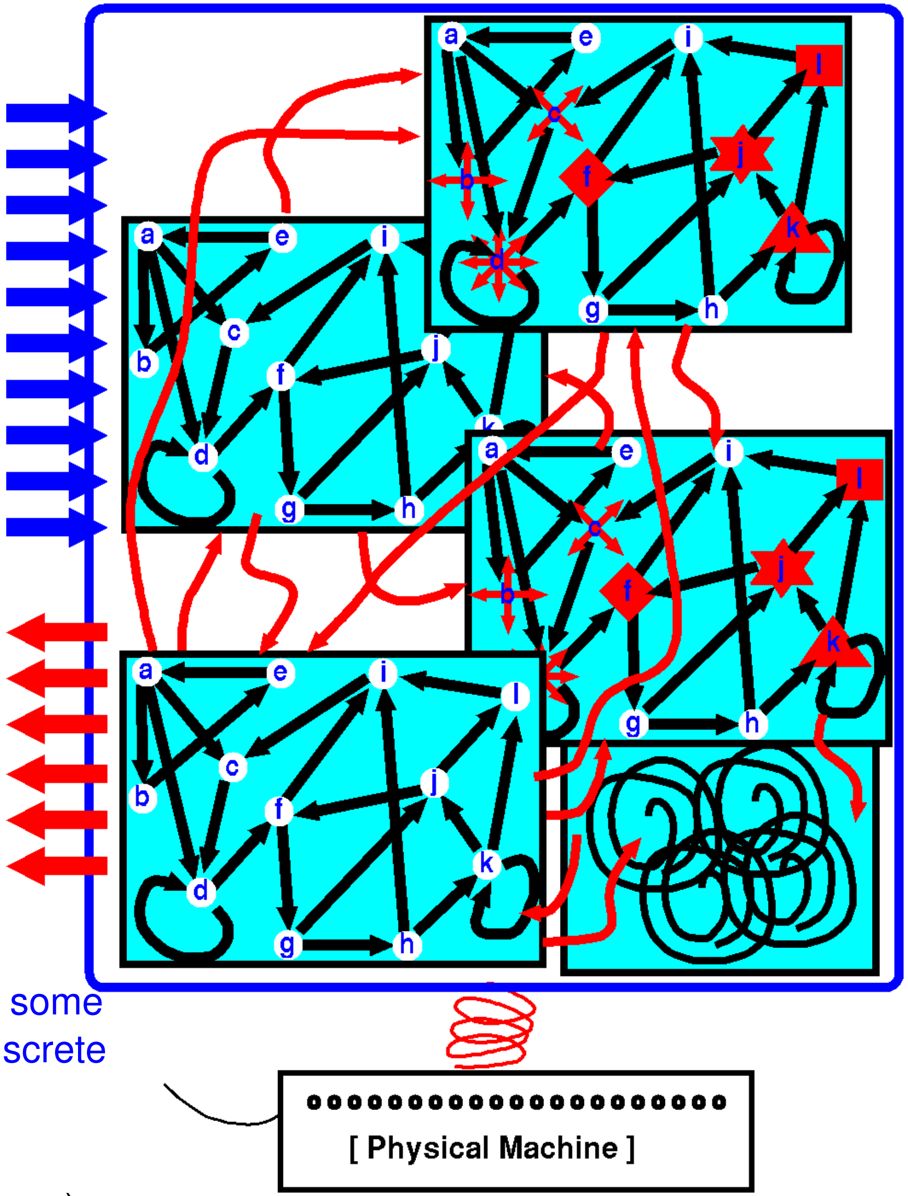
A technical point: software interrupt handlers connected to physical devices that are constantly on, e.g. keyboard and mouse interfaces, video cameras, etc., mean that some processes are constantly “watching” the environment even when they don’t have control of the CPU.
What are the biological analogues?

A more general model

Instead of a fixed set of sub-processes, modern computing systems allow new virtual machine processes to be constructed dynamically,

- of varying complexity
- some of them running for a while then stopping,
- others going on indefinitely.

The red polygons and stars might be subsystems where new, short term or long term, sub-processes can be constructed within a supporting framework of virtual machines – e.g. a new planning process.



If the machine includes analog devices there could be some processes that change continuously, instead of only discrete virtual machines.
Others can simulate continuous change.

(E.g. box with smooth curves, bottom right of VM diagram)

Explaining what's going on in such cases requires a new deep analysis of the notion of **causation**

The relationship between objects, states, events and processes in virtual machines and in underlying implementation machines is a tangled network of causal interactions.

Software engineers have an intuitive understanding of it, but are not good at philosophical analysis.

Philosophers just tend to ignore this when discussing supervenience,

even though most of them use multi-process virtual machines for all their work, nowadays.

Explaining how virtual machines and physical machines are related requires a deep analysis of causation that shows how the same thing can be caused in two very different ways, by causes operating at different levels of abstraction.

Explaining what 'cause' means is one of the hardest problems in philosophy.

For more on the analysis of causation (Humean and Kantian) see:
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#wonac`

Could such virtual machines run on brains?

We know that it can be very hard to control directly all the low level physical processes going on in a complex machine: so it can often be useful to introduce a virtual machine that is much simpler and easier to control.

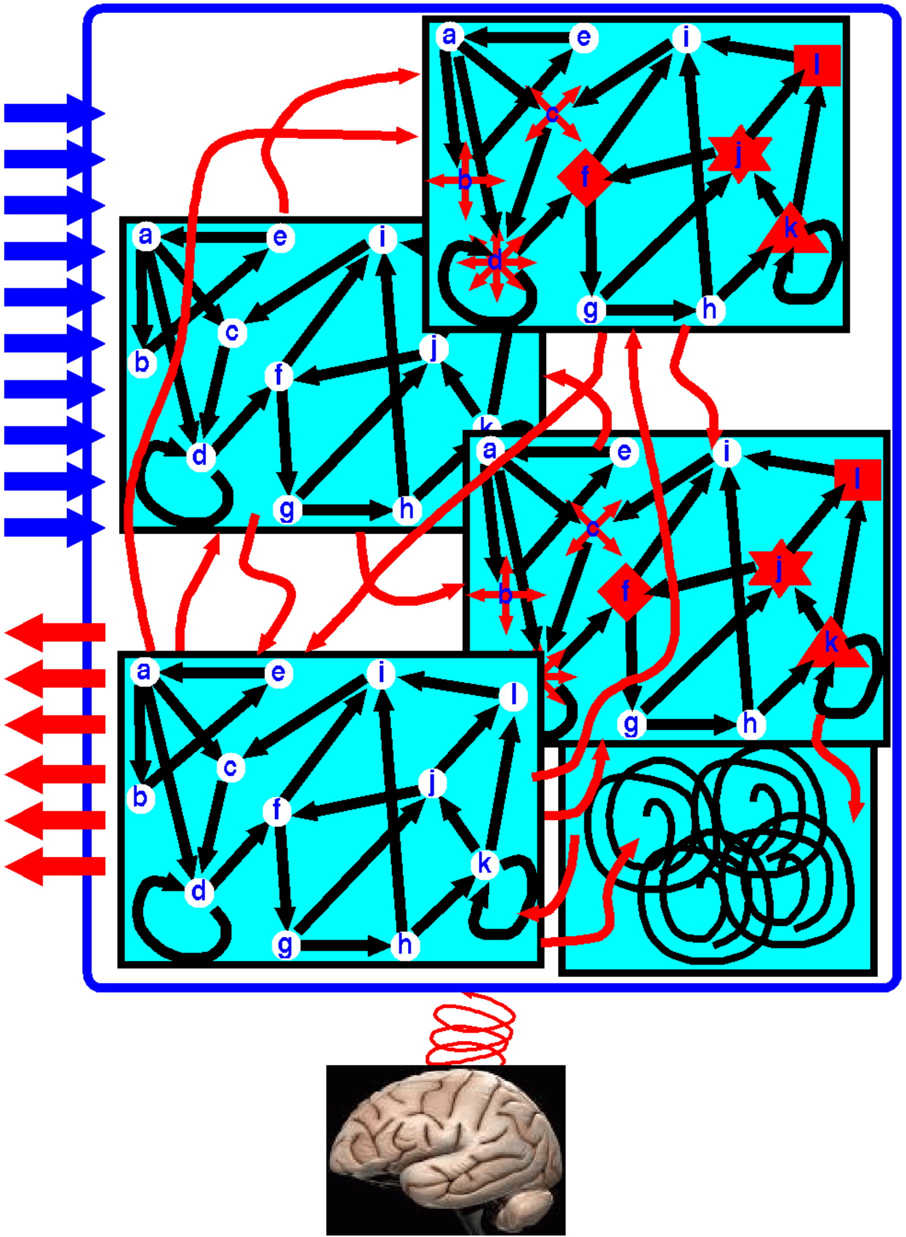
Perhaps evolution discovered the importance of using virtual machines to control very complex systems before we did?

In that case, virtual machines running on brains could provide a high level control interface.

Questions:

How would the genome specify construction of virtual machines?

Could there be things in DNA, or in epigenetic control systems, that we have not yet dreamed of?



Multi-faceted multiprocessing virtual machines

The previous picture might give the impression that all the different concurrently running, constantly interacting, subsystems in a complex system – e.g. an animal mind – are similar in kind, though they may do different things, e.g. having different sets of state-transitions.

But we need to think of some of them as being more remote from the sensorimotor interface to the environment, and doing more abstract things, possibly operating on different time scales.

And many of them may be doing nothing most of the time, lying dormant waiting for something to turn up and wake them up, e.g. a thought or memory may awaken other thoughts or memories, or a sound or something seen may awaken subsystems suited to interpreting the new information and working out what to do about it.

(Resist the trap of thinking that perception is primarily recognition: if it were, you would never be able to see or hear anything new.)

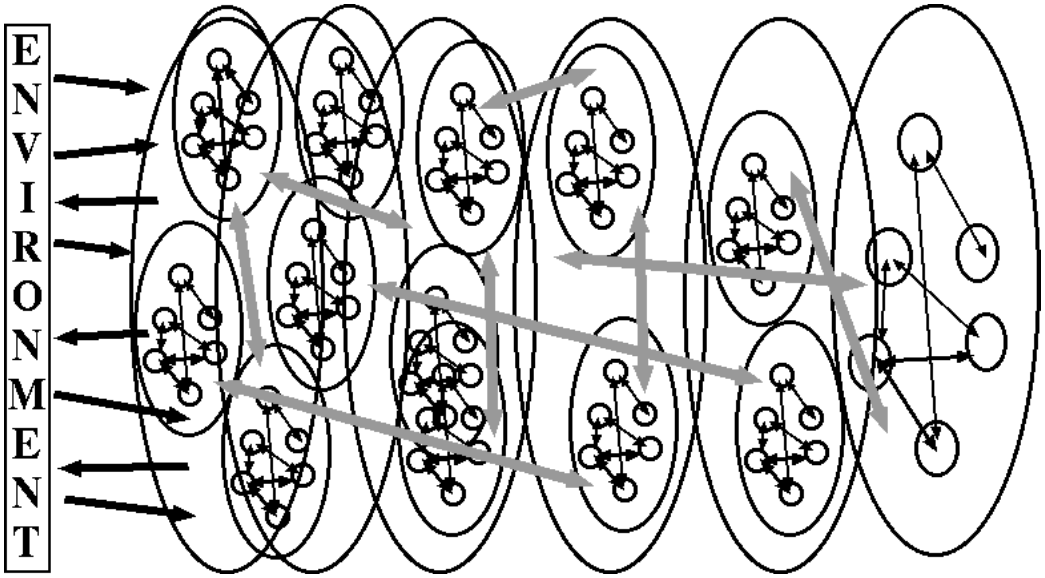
The following slides present an elaborated view of a multi-component biological virtual machine, which need not exist fully formed at birth, but, like a human mind, builds itself over an extended period of developing and learning.

A multi-layered architecture that builds itself

A sophisticated architecture for an intelligent animal or machine includes multiple dynamical systems with many multi-stable components, with many inactive at any time.

Some change continuously, others discretely, and any changes can propagate effects in parallel with many other changes in many directions.

Some components are closely coupled with environment through sensors and effectors, others much less coupled – even free-wheeling, and unconstrained by embodiment (and some of them can use logic, algebra, etc., when planning, imagining, designing, reminisicing, reasoning, inventing stories, etc.).



Most animals have a genetically pre-configured architecture: the human one grows itself, partly in response to what it finds in the environment. Which other animals do that?

For more on this see `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0801a`
Architectural and representational requirements for seeing processes proto-affordances and affordances

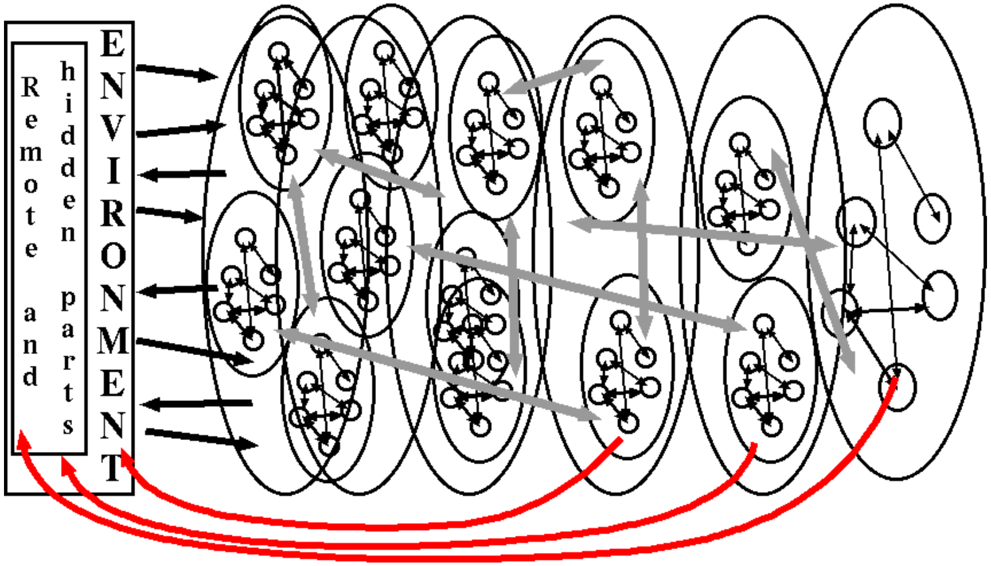
Different parts of the system may have different relationships with the environment

Some portions of such a system may contain information about things going on in the environment, including past, future, and spatially remote events, processes, and individuals.

Think of the red lines as indicating "semantic relations" – i.e. reference to inaccessible parts of the environment.

The portions of the system closest to the "skin" may be primarily concerned with things going on just inside the skin – sensory and motor signals.

They refer only to what is in the body, using only a "somatic" ontology.



Other parts of the system, more detached from sensorimotor business, may use "exosomatic" ontologies, referring to things in the environment:

e.g. they refer to 3-D objects with complex, hidden interiors, to things outside the current house or cave, things that happened in the past, things that may happen in the future, plans for future actions, etc.

That's a major, largely unnoticed, achievement of biological evolution.

How many animals can do that? How did the competence evolve?

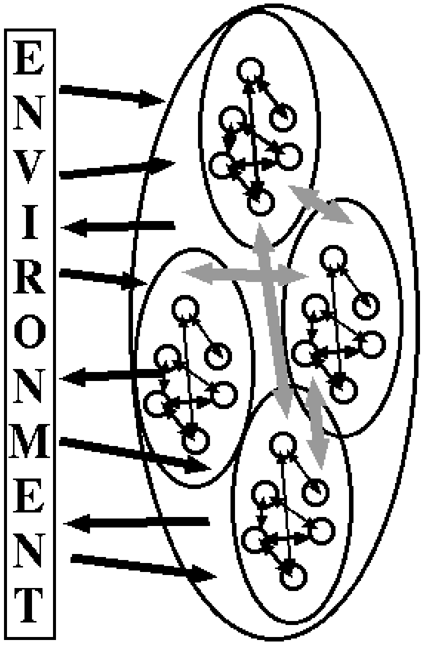
If you think such reference outside the skin is impossible (e.g. because you believe "symbol-grounding" theory), see `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#models`

Contrast the more common view of organisms as dynamical systems

Features of such a system

- All parts of the system are closely coupled with the environment through sensors and effectors.
- Nothing can change internally unless provoked by external changes.
- Everything represented in the system is essentially in the system

  using “somatic” ontologies that refer to patterns in sensor and motor signals
- There is no way to represent possible future processes, distant present events and processes, things in the remote past



That may suffice for microbes, and perhaps even for some insects, but it clearly is not how humans work, and it is very likely that many other animals are not so limited.

Routes from DNA to behaviours: reflexes

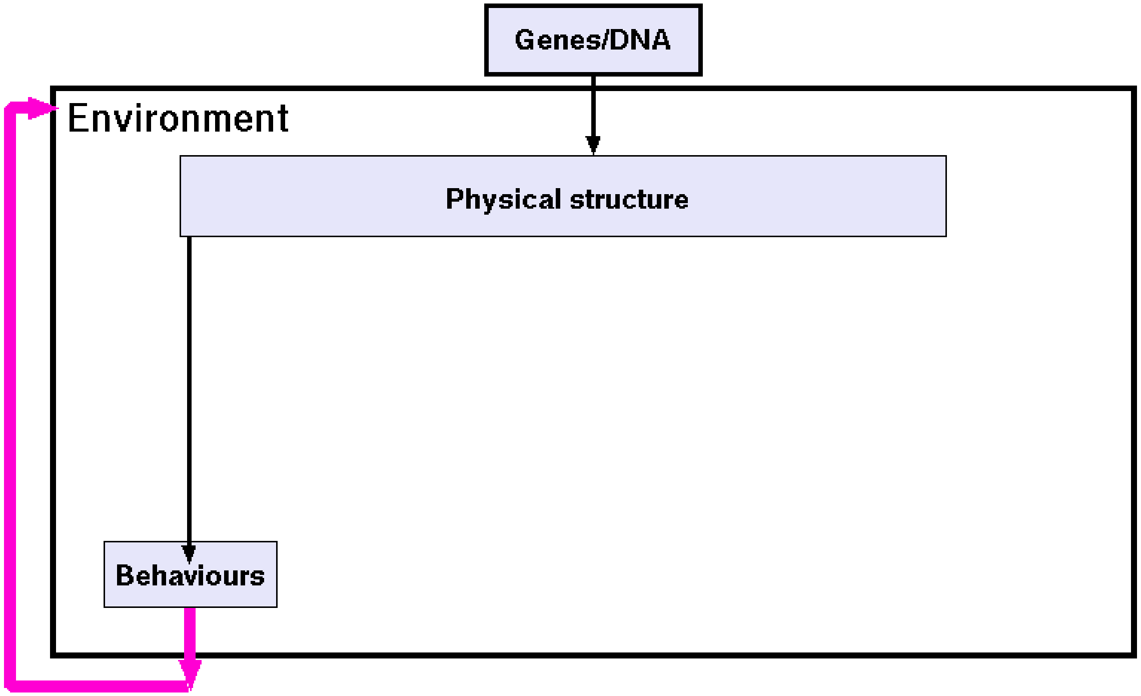
Cognitive epigenesis: Multiple routes from DNA to behaviour, some via the environment

The simplest route from genome to behaviour:

Everything is hard-wired in a design encoded in the genome (subject to interpretation by epigenetic mechanisms).

The physical structures determine

- ongoing behaviours (e.g. breathing, or respiration, or pumping of a heart)
- specific reflex responses to particular stimuli
  e.g. the knee-jerk reflex.



Note: during development, the behaviours that produce effects on the environment may feed back into influencing further development, and learning.

Routes from DNA to behaviours: more flexible competences

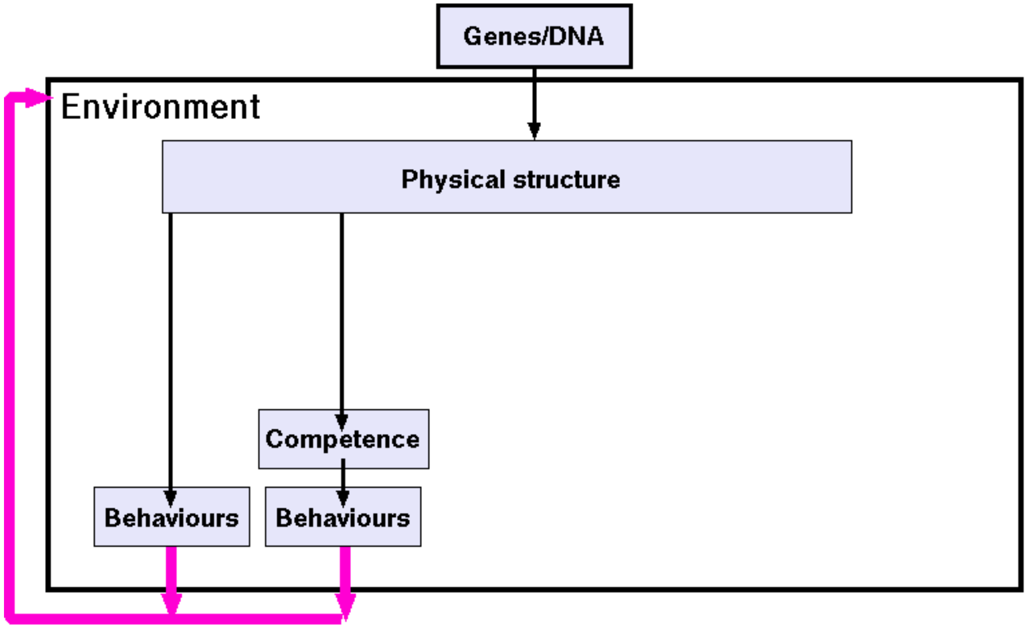
Cognitive epigenesis: Multiple routes from DNA to behaviour, some via the environment

A more complex route from genome to behaviour:

Everything is hard-wired in a design encoded in the genome (subject to interpretation by epigenetic mechanisms).

But what is hard-wired is capable of modifying behaviour on the basis of what is sensed before and during the behaviour.

- The details of such behaviours are products of both the genome and the current state of the the enivonment.



A "competence" is an ability to produce a family of behaviours capable of serving a particular goal or need in varied ways, e.g. picking something up, avoiding an obstacle, getting food from a tree by jumping, catching prey, avoiding predators, migration.

Some competences are "pre-configured" in the genome.

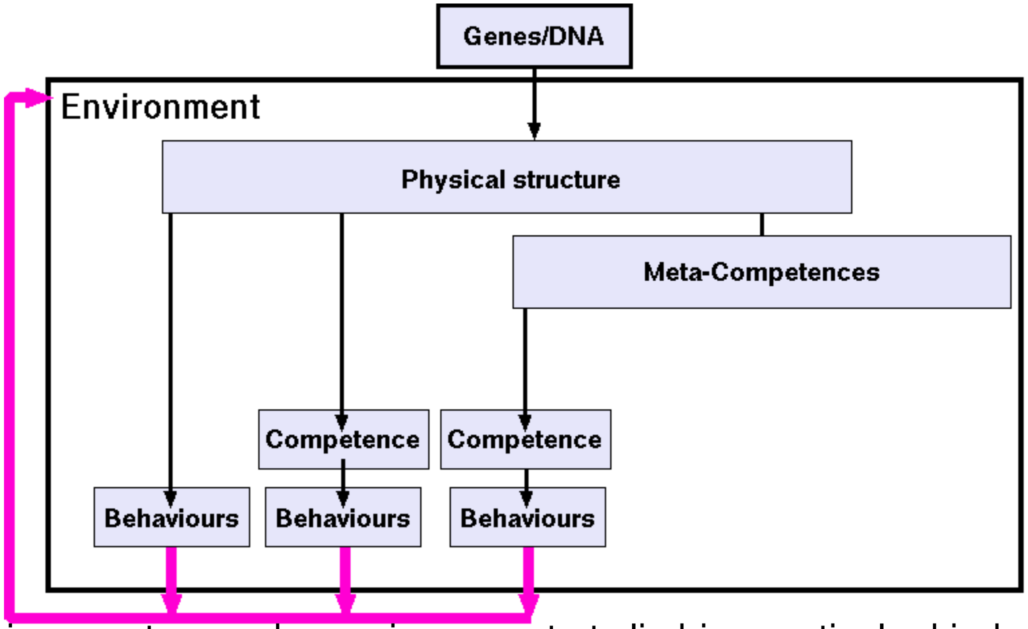
There is not necessarily a sharp division bewteen reflexes and competences: the latter are more flexible and goal directed, but the degree of sophistication can vary a lot.

Routes from DNA to behaviours:
The role of meta-competences

Cognitive epigenesis: Multiple routes from DNA to behaviour, some via the environment

A more complex route from genome to competences:

Instead of competences being hard-wired in a design encoded in the genome (subject to interpretation by epigenetic mechanisms), they may be developed to suit features of the environment, as a result of play, and exploration, leading to learning.
There may be hard-wired genetically preconfigured “meta-competences” that use information gained by experimenting in the environment to generate new competences tailored to the features of the environment – e.g. becoming expert at climbing particular kinds of tree, or catching particular kinds of fish, while conspecifics in another location develop different competences produced by the same mechanisms.



The details of such behaviours are products of both the genome and the current state of the the enivonment.

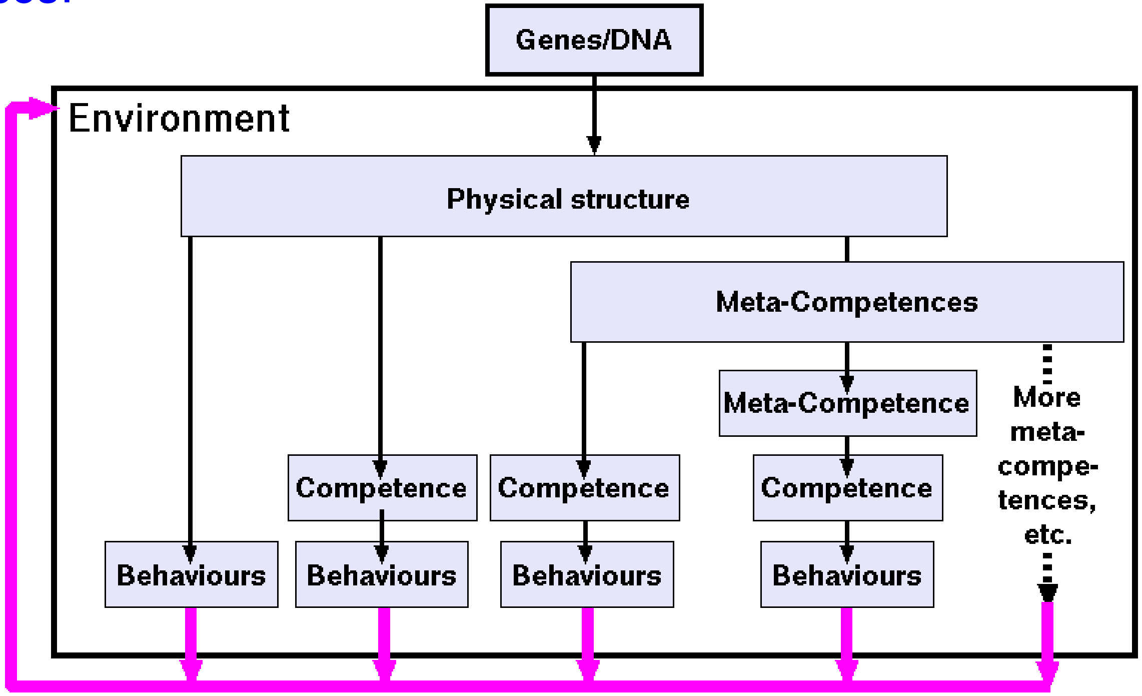
A “meta-competence” is an ability to produce a family of competences capable of serving varied goals in the environment of the animal.

A VM can extend itself indefinitely

Cognitive epigenesis: Multiple routes from DNA to behaviour, some via the environment

Meta-configured meta-competences:

Humans not only learn new kinds of things, they can learn to learn even more varied kinds of things. E.g. after completing a degree in physics yiou are enabled to learn things that a non-physicist could not learn, e.g. learning how to do more sophisticated experiments.



Meta-configured meta-competences:

(towards the right of the diagram) are produced through interaction of pre-configured or previously produced meta-configured competences with the environment, including possibly the social environment

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609`

Natural and artificial meta-configured altricial information-processing systems
Chappell&Sloman, 2007, IJUC

Diagrams developed with Jackie Chappell and Chris Miall.

Self-monitoring and virtual machines

Systems dealing with complex changing circumstances and needs may need to monitor themselves, and use the results of such monitoring in taking high level control decisions.

E.g. which high priority task to select for action.

Using a high level virtual machine as the control interface may make a very complex system much more controllable: only relatively few high level factors are involved in running the system, compared with monitoring and driving every little sub-process, even at the transistor level.

The history of computer science and software engineering since around 1950 shows how human engineers introduced more and more abstract and powerful virtual machines to help them design, implement, test debug, and run very complex systems.

When this happens the human designers of high level systems need to know less and less about the details of what happens when their programs run.

Making sure that high level designs produce appropriate low level processes is a separate task, e.g. for people writing compilers, device drivers, etc. Perhaps evolution produced a similar "division of labour"?

Similarly, biological virtual machines monitoring themselves would be aware of only a tiny subset of what is really going on and would have over-simplified information.

THAT CAN LEAD TO DISASTERS, BUT MOSTLY DOES NOT.

Robot philosophers

These inevitable over-simplifications in self-monitoring could lead robot-philosophers to produce confused philosophical theories about the mind-body relationship.

Intelligent robots will start thinking about these issues.

As science fiction writers have already pointed out, they may become as muddled as human philosophers.

So to protect our future robots from muddled thinking, we may have to teach them philosophy!

BUT WE HAD BETTER DEVELOP GOOD PHILOSOPHICAL THEORIES FIRST!

The proposal that a virtual machine is used as part of the control system goes further than the suggestion that a robot builds a high level model of itself, e.g. as proposed by Owen Holland in

`http://cswww.essex.ac.uk/staff/owen/adventure.ppt`

For more on robots becoming philosophers of different sorts see

Why Some Machines May Need Qualia and How They Can Have Them:
Including a Demanding New Turing Test for Robot Philosophers

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0705`
Paper for AAAI Fall Symposium, Washington, 2007

AI Theorists make philosophical mistakes

A well known “hypothesis” formulated by two leading AI theorists, Allen Newell and Herbert Simon is The Physical Symbol System Hypothesis, stating that

A physical symbol system has the necessary and sufficient means for intelligent action.

They state that a physical symbol system “consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure). Thus, a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another)....”

See `http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/PhysicalSymbolSystemHyp.html`

It should be clear to anyone who is familiar with how AI programming languages work that there is a deep flaw in this: the symbols manipulated by AI systems are not physical objects or even physical patterns: they are abstract objects that inhabit virtual machines, but are implemented in physical machines.

E.g. a bit pattern in a computer memory is not the same thing as the physical state of a collection of transistors, since the actual correspondence between bit patterns and physical details is quite complex, and may be different in different parts of the same computer (e.g. in different types of memory used and in the CPU, especially where memory uses redundant self-correcting mechanisms).

Moreover the most important relations between bit patterns do not involve physical proximity but locations in a virtual address space – e.g. one bit pattern can encode the address of another and adjacency in the virtual address space is what matters, not physical adjacency.

Instead of a physical symbol system they should have referred to a Physically Implemented Symbol System. Perhaps they did not wish to refer to a PISS??

Coping with novelty

The history of human science, technology and art shows that people are constantly creating new things – pushing the limits of their current achievements.

Dealing with novel problems and situations requires different mechanisms that support creative development of novel solutions.

(Many jokes depend on that.)

If the deeper, more general, slower, competence is not available when stored patterns are inadequate, wrong extrapolations can be made, inappropriate matches will not be recognised, new situations cannot be dealt with properly and further learning will be very limited, or at least very slow.

In humans, and probably some other animals, the two systems work together to provide a combination of fluency and generality. (Not just in linguistic competence, but in many other domains.)

Where does the human power come from?

Before human toddlers learn to talk they have already acquired deep, reusable structural information about their environment and about how people work.

They cannot talk but they can see, plan, be puzzled, want things, and act purposefully: They have something to communicate about.

E.g. see Warneken's videos.

That pre-linguistic competence grows faster with the aid of language, but must be based on a prior, internal, formal 'linguistic' competence

using forms of representation with structural variability and (context-sensitive) compositional semantics.

This enables them to learn any human language and to develop in many cultures.

Jackie Chappell and I have been calling these internal forms of representation Generalised Languages (GLs).

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang`
What evolved first: Languages for communicating, or languages for thinking
(Generalised Languages: GLs)

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0703`
Computational Cognitive Epigenetics

More papers/presentations are on my web site.

Can we give robots suitable GLs ?

Robots without a similar pre-communicative form of representation allowing them to represent things about their world, and also possible goals to achieve, will not have anything to communicate when they start learning a language.

Their communicative competences are likely to remain shallow, brittle and dependent on pre-learnt patterns or rules for every task because they don't share our knowledge of the world we are talking about, thinking about, and acting in.

Perhaps, like humans (and some other altricial species), robots can escape these limitations if they start with a partly 'genetically' determined collection of meta-competences that continually drive the acquisition of new competences building on previous knowledge and previous competences: a process that continues throughout life.

The biologically general mechanisms that enable humans to grow up in a very wide variety of environments, are part of what enable us to learn about, think about, and deal with novel situations throughout life.

I conjecture that this requires an architecture that grows itself over many years partly as a result of finding out both very general and very specific features of the environment through creative play, exploration and opportunistic learning.

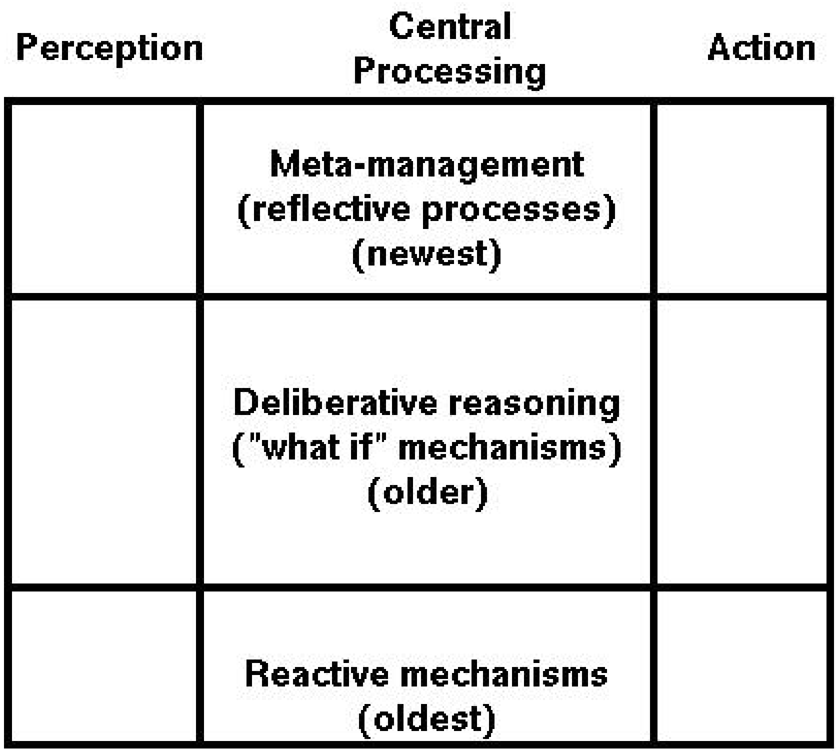
Development of brains of some animals has to be staggered so that some parts don't start learning until others have developed competences from which new things can be learnt.

We can discern some major sub-divisions within a complex architecture

The CogAff Schema – for designs or requirements for architectures.

Different layers correspond to different evolutionary phases.

NEWER



Perception	Central Processing	Action
	Meta-management (reflective processes) (newest)	
	Deliberative reasoning ("what if" mechanisms) (older)	
	Reactive mechanisms (oldest)	

OLDER

(Think of the previous “dynamical systems” diagram as rotated 90 degrees counter-clockwise.)

H-Cogaff: the human variety

Here’s another view of the architecture that builds itself (rotated 90 degrees).

This crudely indicates a possible way of filling in the CogAff schema on previous slide – to produce a human-like architecture (H-CogAff).

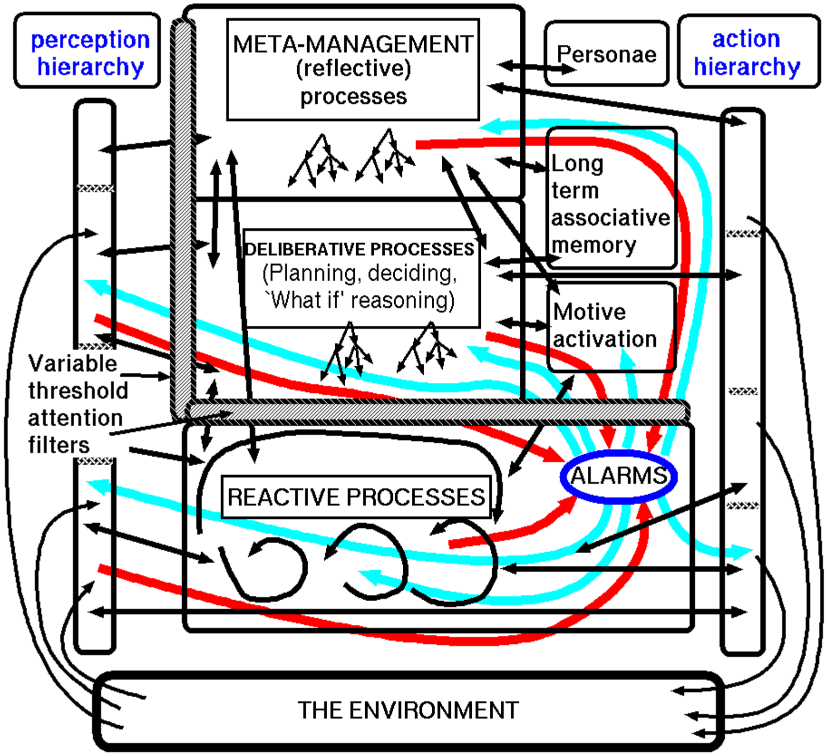
NEWER

The higher level capabilities evolved much later than the others.

(Arrows represent flow of information and control)

OLDER

The reactive behaviours are likely to be older.


perception hierarchy
META-MANAGEMENT (reflective) processes
Personae
action hierarchy
Long term associative memory
DELIBERATIVE PROCESSES (Planning, deciding, `What if' reasoning)
Motive activation
Variable threshold attention filters
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

For more details see papers and presentations in the Cogaff web site:

`http://www.cs.bham.ac.uk/research/projects/cogaff/`

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks`

A PICTURE OF YOUR MIND?

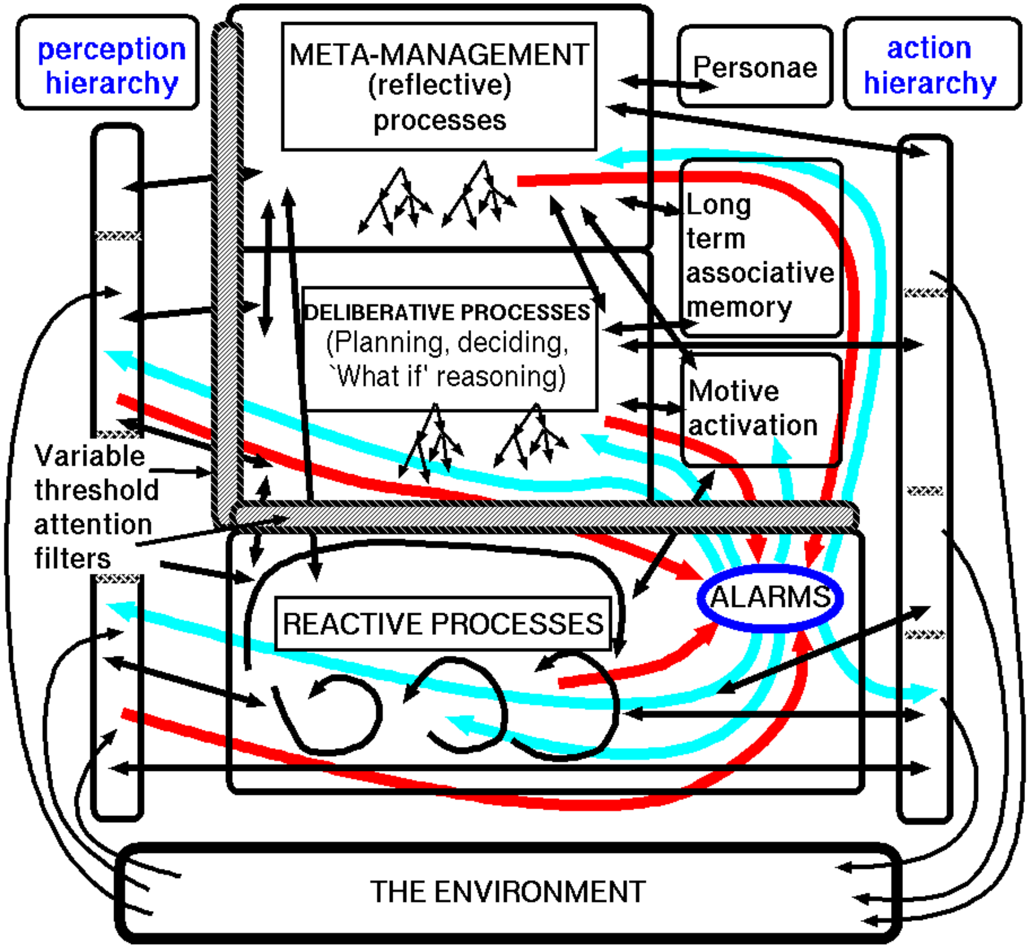
What sort of virtual machine runs on your brain? Here's a crude picture: The H-CogAff architecture.

The Birmingham Cognition and Affect project proposed a general schema (CogAff) for architectures, including ancient biological reactive mechanisms (including "alarm" systems), less ancient biological deliberative mechanisms (e.g. for making long term predictions, future plans, and explaining things) and even newer "metamanagement" mechanisms for self-monitoring and self-control.

The CogAff schema seems to cover many kinds of designs, ranging from very small and simple organisms to more complex designs.

A special case of the CogAff schema is the H-CogAff (Human-CogAff) architecture, shown crudely here.

So far only small parts of this have been implemented.


perception hierarchy
META-MANAGEMENT (reflective) processes
Personae
action hierarchy
Long term associative memory
DELIBERATIVE PROCESSES (Planning, deciding, `What if' reasoning)
Motive activation
Variable threshold attention filters
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

See also: the presentations on architectures here:
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/`

What would a robot with the H-CogAff VM Architecture be like?

- It would have a lot of innate or highly trained reactive behaviours.
- It would be able to do some planning, explaining, predicting, hypothesising, designing, story telling, using its deliberative mechanisms.
- Its metamanagement methods examining and controlling the robot's own high level virtual machine, as well as perhaps thinking about and communicating with others, would probably under some circumstances start doing philosophical speculation about the nature of its own mind.
- The result will probably be a lot of deep philosophical confusion.
- Unless we can teach it to be a good philosopher.
- For a start, we could ask it to study and analyse these slides and evaluate them as presenting a theory about how the robot works.
- Maybe some of them will come up with much better philosophical theories about minds and bodies than any human philosophers have done.

Machines that have their own motives

Human learning is based in part on what humans care about
and vice versa: what humans care about is based in part on what they learn.

In particular, learning to help others depends on caring about what they want, what will and will not harm them, etc.

And learning about the needs and desires of others can change what we care about.

The factors on which the caring and learning are based can change over time – e.g. as individuals develop or lose capabilities, acquire new preferences and dislikes, etc.

If we make machines that can come to care, then that means they have their own desires, preferences, hopes, fears, etc.

In that case, we shall have a moral obligation to take account of their desires and preferences: anything with desires should have rights: treating them as slaves would be highly immoral.

As noted in the Epilogue to *The Computer Revolution in Philosophy* (1978), now online here:

`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`

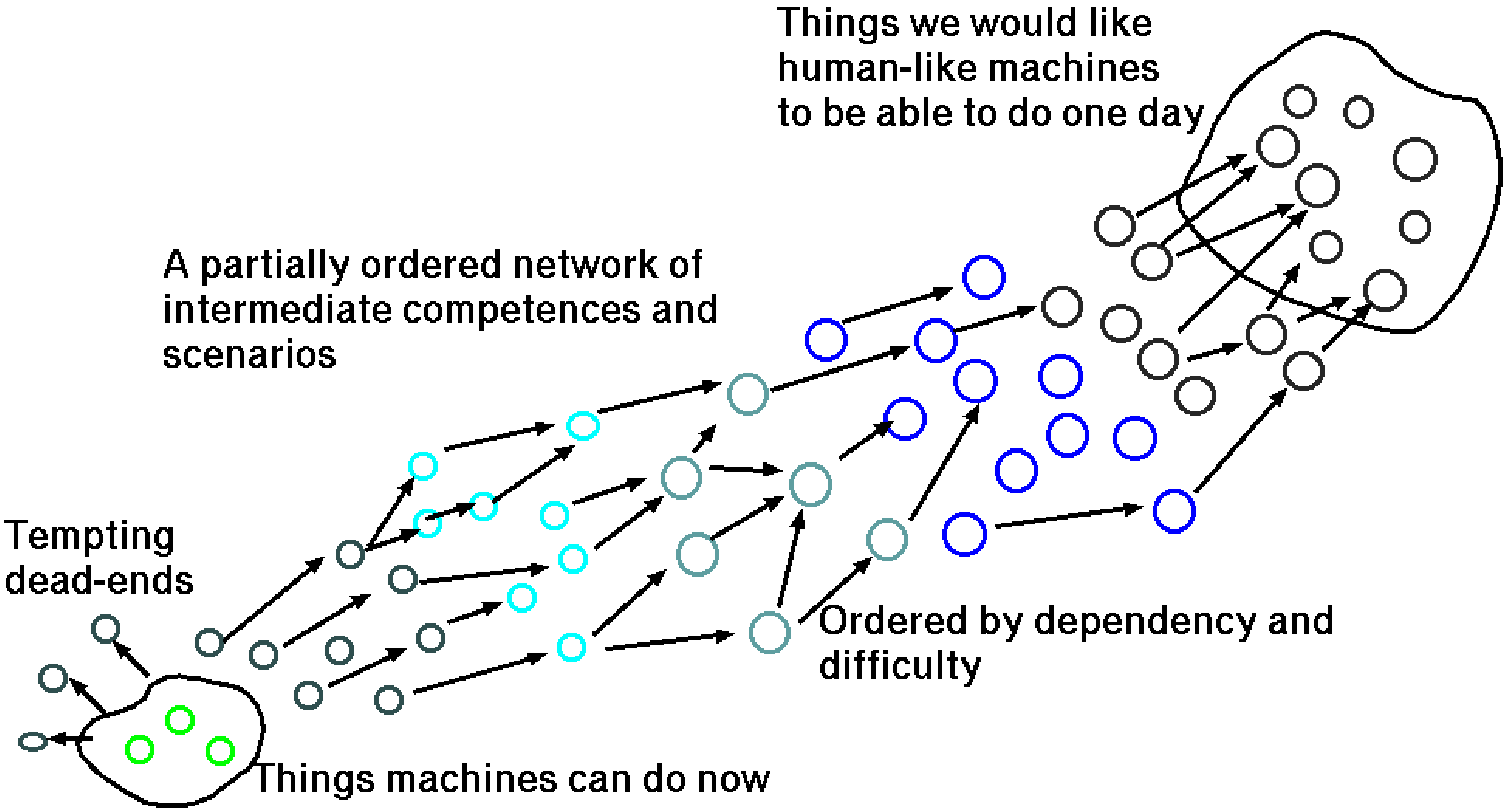
And in this paper on why Asimov's laws of robotics are immoral:

`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/asimov-three-laws.html`

Can it be done?

Perhaps we should work on this roadmap as a way of understanding requirements?

We can take care to avoid the stunted roadmap problem.



Things we would like
human-like machines
to be able to do one day

A partially ordered network of
intermediate competences and
scenarios

Tempting
dead-ends

Ordered by dependency and
difficulty

Things machines can do now

THANK YOU!

For a lot more on supervenience and virtual machines see

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#super`

For ideas about how machines or animals can use symbols to refer to unobservable entities see

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#models`

Introduction to key ideas of semantic models, implicit definitions and symbol tethering

For an argument that internal generalised languages (GLs) preceded use of external languages for communication, both in evolution and in development, see

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang`

What evolved first: Languages for communicating, or languages for thinking
(Generalised Languages: GLs) ?

Additional papers and presentations

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/`
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/`

See also the URLs on earlier slides.