SILICON SOULS

HOW TO DESIGN A FUNCTIONING MIND

Inaugural lecture, 18 May 1992:

The University of Birmingham

Aaron Sloman

- **A brief introduction to Artificial Intelligence (the "core" of Cognitive Science)**

- **Based on the simple idea that a mind is a sophisticated self-modifying control system.**

- **This idea, when developed, has profound implications, (for philosophy, psychology, education, counselling...).**

- **It is not provable or refutable: it defines an approach to the study of mind.**

- **It is not possible to argue against those who believe minds include a "magical" element inexplicable by scientific (mechanistic) theories of mind: the issue is not rationally discussable. I shall simply ignore it.**

- **Some objections rely on inadequate concepts of "mechanism". I'll try to outline the new, broader, concept of mechanism inspired by Computer Science: a mind-stretching exercise. (In one hour????)**

WHY IS IT EASIER FOR COMPUTERS TO GUIDE A ROCKET TO THE MOON THAN TO SIMULATE A HUMAN CHILD, OR EVEN A SQUIRREL?

PLAN OF THIS TALK

1 What is AI?
   - The general study of intelligent systems (badly named)
2 Types of scepticism about AI.
3 Approaches to the study of mind. Choose the “design-based” approach: philosophy as engineering design.
4 Types of control systems. Start with some very simple ones: thermostats. Then elaborate.
5 What’s special about Intelligent control systems?
   - sharing channels between control functions
   - many layers of interpretation and decoding
   - many layers of control
   - multiple independently variable, interacting, sub-states
   - rich functional differentiation
   - structural variability, not just quantitative variation
   - internal self-monitoring, and self-modification
   - and more .....
6 What sort of underlying engine is needed?
7 The space of possible designs: the “shape” of design space.
8 Some conjectures
9 If there’s time: Prospects at Birmingham
10 Summary: Types of optimism about AI.

1. WHAT IS ARTIFICIAL INTELLIGENCE?

Partial and misleading definitions abound. Errors include:

- **Restricting AI to expert systems (a subset of Applied AI)**
- **Restricting AI techniques to logical inference (just one of many forms of computation)**
- **Restricting AI to a branch of applied science (its goals are to understand and explain, not just to make things)**
- **Restricting AI to what can be done using current computer know-how (we know very little, as yet).**
- **Restricting applied AI to hardware and software design.**

Towards a broader view:

Look at AI journals, at AI centres in industry and academe, at conferences, at books. We find that AI is:

- **Multi-disciplinary: philosophy, psychology, linguistics, anthropology, neuroscience and computer science.**
- **Based on computing developments that transform our ideas about mechanisms (e.g. “virtual machines”).**
- **A long term project (so far it’s preliminary exploration, despite claims of high priests of each new fashion)**
- **Aimed at understanding not only human intelligence, but also various kinds of animal intelligence and artificial intelligences: i.e. it’s a general study of possible types of mind (or behaving system).**
- **Potentially able to give us new insights into “affective” (motivational and emotional) aspects of mind, with applications in counselling and therapy.**
- **Plagued with myriad HARD, UNSOLVED problems: it’s in its infancy still. We need bright people to join the field!**

2. TYPES OF PESSIMISTS ABOUT AI

Three types of pessimists about the long-term prospects for AI:

1. **Wishful thinkers:** **Those who wish it to fail, from fear of consequences, or worse: loss of self-importance. Compare hoping to keep humans at the centre of the Universe, or unique among animals. Wishful thinking didn't help there either.**
   (If talk of souls isn't just empty waffle, AI may show us how to create them in laboratories: a stronger challenge to most forms of theology than Copernicus, Darwin, or cosmology. Most clerics haven't noticed!)
2. **Ignorant, unimaginative, or mystics:** **Those who are ignorant and can't see how AI could possibly succeed: their view of mind may be too mystical or their concept of computation, or of mechanism, too limited. E.g.**
   -- "computers do only what they are told to do" (but not self-programming computers).
   -- "I can't imagine a machine being creative, having emotions, etc."
   (But what you can't imagine merely shows your limitations: compare space-filling curves, wave-like particles, cars moving with increasing acceleration and decreasing velocity, etc.)
3. **Informed pessimists:** **Those with detailed knowledge who can see why it is so difficult even to get machines to do what young children or squirrels can do. (Playing chess, or solving mathematical problems is much easier and computers already outperform most of us!) Sometimes these people produce arguments that help to define or clarify the tasks of AI! (E.g. Dreyfus)**

I'll discuss types of optimists later.

“OBVIOUS” RESPONSES TO SCEPTICISM

Show impressive videos and list achievements of AI, e.g.

- **Expert systems (diagnosing, advising, checking, planning)**
- **Robots of various kinds (mostly very rigid and limited). E.g.: Hopping machines, robot assemblers**
- **Vision systems (e.g. quality control, robot control)**
- **Natural language front ends (all restricted)**
- **“Intelligent” tutoring systems**
- **Chess machines**
- **Aids to mathematical problem solving**
- **Neural nets**
- **Powerful software development tools and AI languages (which could transform many kinds of programming)**

UNFORTUNATELY:

All the examples are miles away from explaining even the abilities of a child, or chimp, or squirrel.

Instead of a catalogue of (not always very impressive) achievements I’ll address general issues.

In order to make real progress in understanding we need a far deeper grasp of:

(a) what intelligent systems need to be able to do, and
(b) what various kinds of mechanisms can do.

BUT:

OUR UNDERSTANDING IS STILL VERY SHALLOW

E.g. our grasp of mechanism is shallow: most people don’t know what computation is. Experts disagree too.

3. APPROACHES TO THE STUDY OF MIND

1. Species-based / biology-based approaches:

Study and try to model and understand some existing intelligent systems (usually humans). Examples:

1.a. Semantics-based (try to explain ordinary use of mental concepts)

1.b. Phenomena-based (look for correlations between phenomena, assuming you know what you are talking about: e.g. study causes and effects of "joy", "hate", etc.).

2. Mechanism-based approaches (bottom up):

Take a particular class of mechanisms (computers, symbol processors, neural nets, etc.) and explore what can be done with them. (Some people in AI, including most connectionists, work like this.)

3. Design-based approach:

Explore the "space of possible designs" (mechanisms and architectures)

(a) Both known and unknown mechanisms

(b) Existing and merely possible "species" etc.

(c) Top down and bottom up approaches combined

(Different approaches need to be combined)

Studying a design requires more than building a working system: It requires understanding which features are important for which capabilities, and how the capabilities would change if the design were changed, etc. Cognitive science needs this kind of "design stance".

4. Philosophy: try to "deduce" the only possible design.

Engineers know solutions aren't unique: there are always trade-offs: Philosophers should be engineers. (Some are).

4. THE MIND (OR BRAIN) AS A CONTROL SYSTEM

There are many different ways of thinking about the mind. At a certain level of abstraction we can think of it as:

an incredibly complex, self-monitoring, self-modifying control system

How then is it like and how is it unlike other control systems?

There's a large body of mathematics concerning control systems. Does it help us understand how minds work?

ANSWER: "not much"!

THIS IS BECAUSE:

- **The architecture is so rich: there's enormous functional differentiation within each individual.**
- **The architecture is not static, it develops over time. (So a fixed set of differential equations can't model it).**
- **The most important changes and processes don't map onto numeric variation: many are structural.**

So that:

- **Causal influences are not all expressible as transmission of measurable quantities like force, current, etc. Some transmit structured "messages" and instructions. Some build new structures (embryoes).**
- **New kinds of mathematics are needed to cope with this.**
- **Abstract or "virtual" machines, implemented in terms of lower level physical machines, manipulate complex information structures (e.g. networks of symbols) rather than physical objects and their physical properties.**
  E.g. a word-processor manipulates words, paragraphs, etc.

SOME KEY IDEAS

We need new thinking tools to help us grasp all this complexity. AI has provided many detailed concepts for thinking about and modelling processes in perception, planning, reasoning, learning. I want to talk about more "global ideas" to help us think about the architecture of the whole mind: how the bits fit together. It's all still a bit vague, and in need of development. Key notion: sub-state.

ATOMIC VS MOLECULAR STATES AND SUB-STATES

Contrast two different concepts of state change, the first familiar from physics and engineering ("state space"):

- **atomic state: The whole system state is thought of as indivisible, and the system moves from one state to another through a "state space" or "phase space".**
- **molecular state with sub-states: The instantaneous state includes many coexisting, independently variable, interacting, states of different kinds, which change in different ways, under external or internal influences.**

TOWARDS A THEORY OF INTERACTING SUB-STATES

Some control systems include:

- **Desire-like control states (associated with "attractors"†)**
- **Belief-like control states**
- **Loose and indirect causal links between input and output channels and the control states**
- **Time-sharing of causal channels**
- **Structural, and not only quantitative, change**

I'll try to draw out some of the implications of all this.

† (Thanks to George Kiss at the Open University for this analogy from dynamical systems theory. It's only partial, since a desire-like, but inactive, state need not actually generate behaviour)

CONTROL SYSTEM ARCHITECTURES

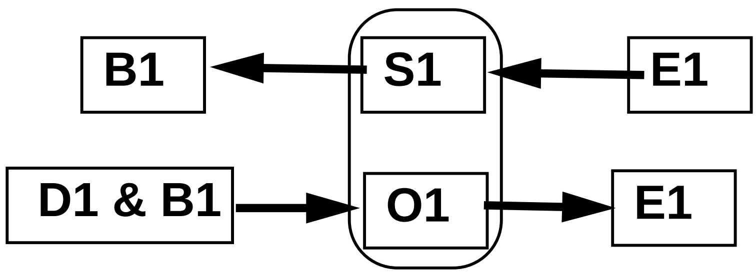
Systems vary in their underlying mechanisms (e.g. neural, symbolic, etc.), and, more importantly, in their architectures. Different (changeable) control sub-states may have different functional roles.

Control sub-states vary independently: variation is one- or N-dimensional, structural, continuous, discrete, etc.

ARCHITECTURE OF A THERMOSTAT (simplified):

A thermostat with a temperature sensor and a control knob has two control states, one belief-like (B1) set by the sensor and one desire-like (D1), set by the knob.

- **B1 tends to be modified by changes in a feature of the environment E1 (its temperature), using an appropriate sensor (S1), e.g. a bi-metallic strip.**
- **D1 tends, in combination with B1, to produce changes in E1, via an appropriate output channel (O1)**



This is a particularly simple feedback control loop:

The states (D1 and B1) both admit one-dimensional continuous variation. D1 is changed by “users”, e.g. via a knob or slider, not shown in this loop.

Other architectures differ in the kinds of sub-states, the number and variety of sub-states, the functional differentiation of sub-states, the kinds of causal influences, etc. E.g. could a machine change its own D1?

A MULTI-CHANNEL CONTROL SYSTEM

Systems with more complex architectures simultaneously control several different aspects of the environment,

E1, E2, E3, etc.

using sensors: S1, S2, S3, etc.

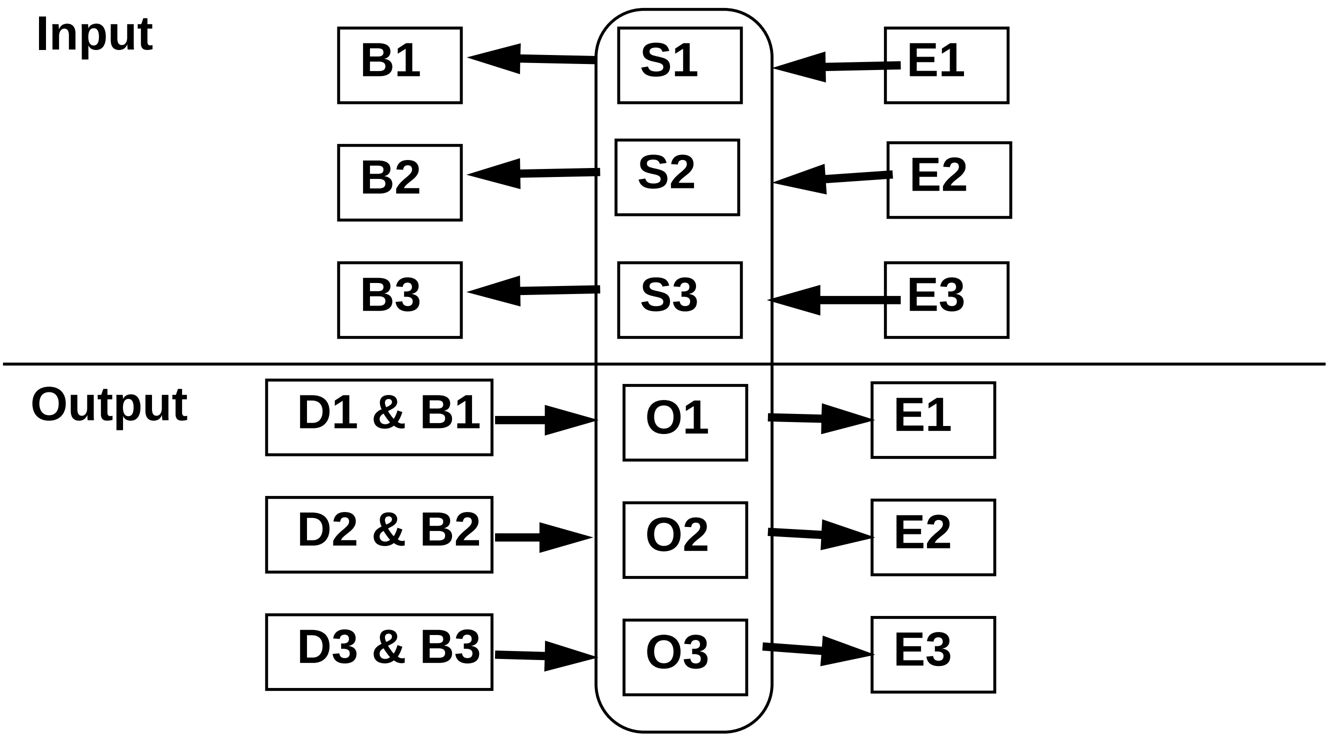
and output channels: O1, O2, O3, etc.

which are causally linked to belief-like internal states:

B1, B2, B3, etc,

and desire-like internal states:

D1, D2, D3 etc.

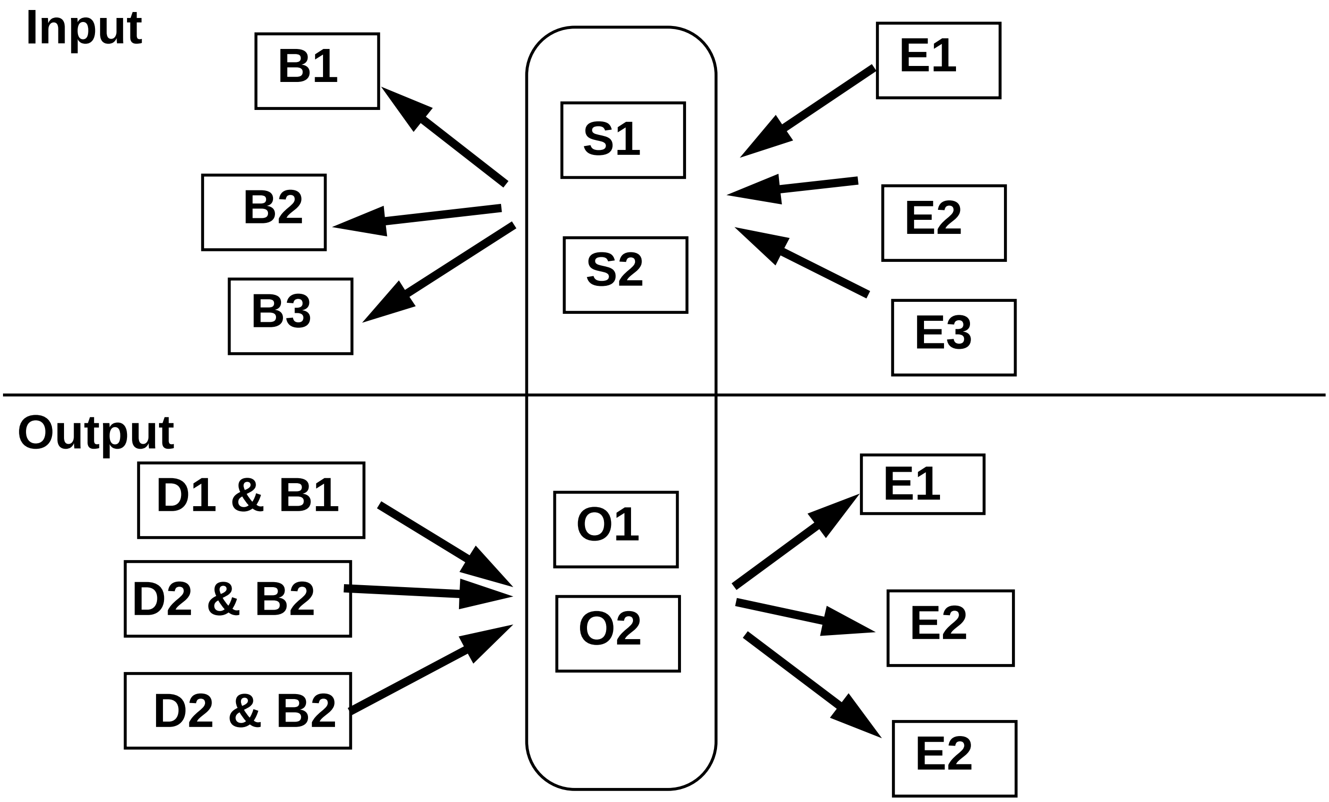


Essentially this is just a collection of separate feedback loops.

The architecture can be more complicated in various ways: e.g. sharing channels, layers of input or output processing, self monitoring, self-modification, etc..

SHARED INPUT AND OUTPUT CHANNELS

Instead of having separate sensors (Si) and output channels (Oi) for each Environmental property, belief-like and desire-like state (Ei, Bi, Di) a complex system might time-share a collection of Si and Oi between different sets of Ei, Bi, Di, e.g.



EXAMPLES

- **<u>Sharing two eyes</u> (S1, S2) between a collection of beliefs about different bits of the environment**
- **<u>Sharing two hands</u> (O1, O2) between different desires relating to the state of the environment.**

Or, at a lower level: sharing millions of visual pathways, and millions of motor pathways among a smaller (?) collection of beliefs and desires.

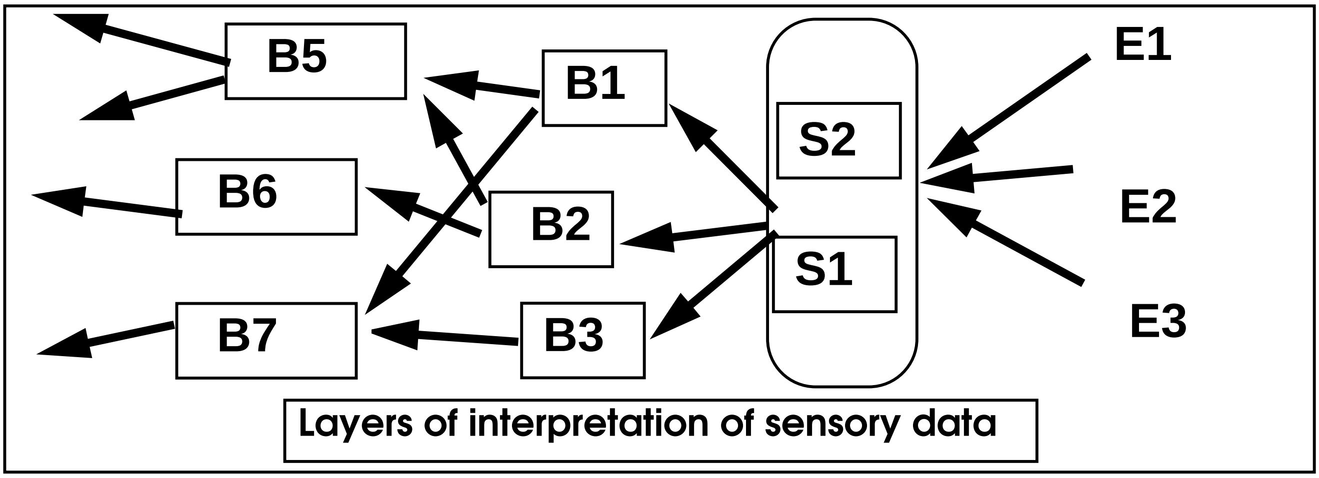
Sharing may be simultaneous or serial.

FURTHER COMPLICATIONS OF DESIGN (1)

BELIEF-LIKE SUB-STATES (Bi)

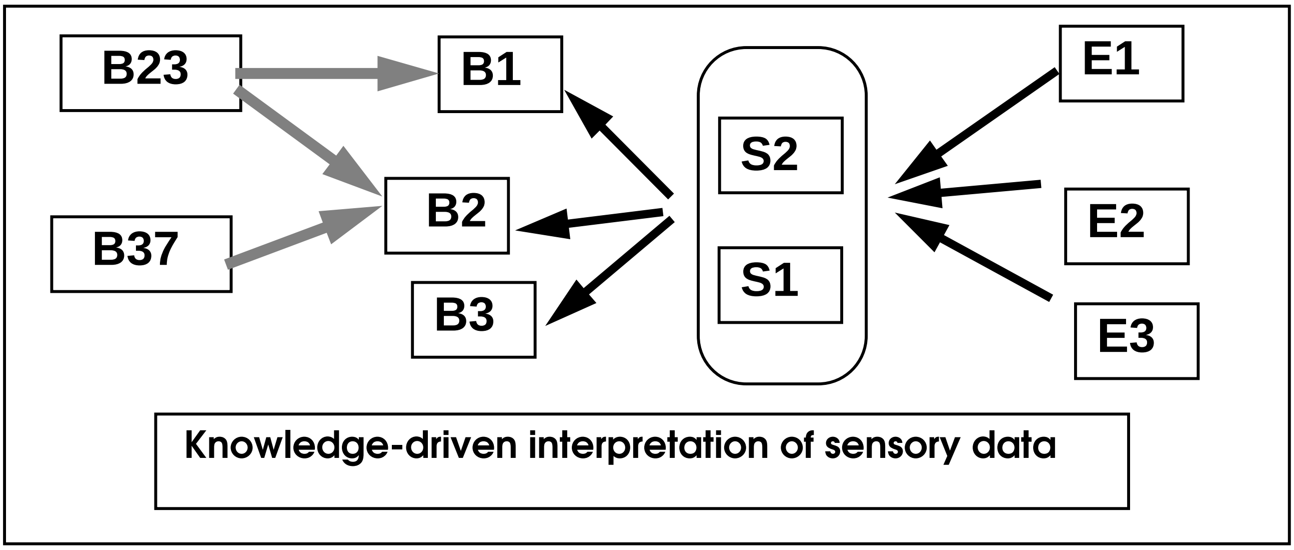
Production of belief-like states can be more complicated:

- **Sharing input channels between different Ei and Bi necessitates <u>interpretation</u> processes, to extract information relevant to different Bi from sensory “arrays”. (Normally this requires specialised knowledge: <u>general</u> principles do not suffice for disambiguation. E.g. getting 3-D structure from 2-D visual arrays: a mathematically indeterminate problem.)**
- **<u>Many layers of interpretation</u>: different depths of processing of incoming information. (E.g. phonemes, words, phrases, meanings, theories.)**



- **Different layers of interpretation may use different <u>forms</u> of information storage: retino-topic, analogical, histograms, “structural descriptions” (e.g. trees, networks), labels for recognised complexes, etc. Shape representation is an unsolved problem.**
- **Different intermediate “databases” may be used for different purposes. (E.g. posture control vs recognition.)**

- Some or all of the Bi may be produced or modified on the basis not only of incoming information, but also using previously stored information (e.g. knowledge-driven, partly “top-down” perception).



- Some of the Bi may be stored for future use, or may modify previous long term information stores. Some Bi will be generalisations of many particular Bi.
- Internal self monitoring is possible: some control loops involve only internal processes and substates. Then the Bi record Ei that are internal states: not all monitoring is of the environment. (Steps towards self-consciousness.)
- Time-sharing of input channels may require inputs received at different times to be integrated for certain of the Bi. (E.g. looking at different parts of a house in order to grasp its structure). Implications for storage.

ALL OF THESE POINTS HAVE IMPLICATIONS FOR THE ARCHITECTURE (THE GLOBAL DESIGN) OF A PERCEIVING AGENT

PERCEPTUAL ARCHITECTURES

E.g. visual mechanisms need to take account of:

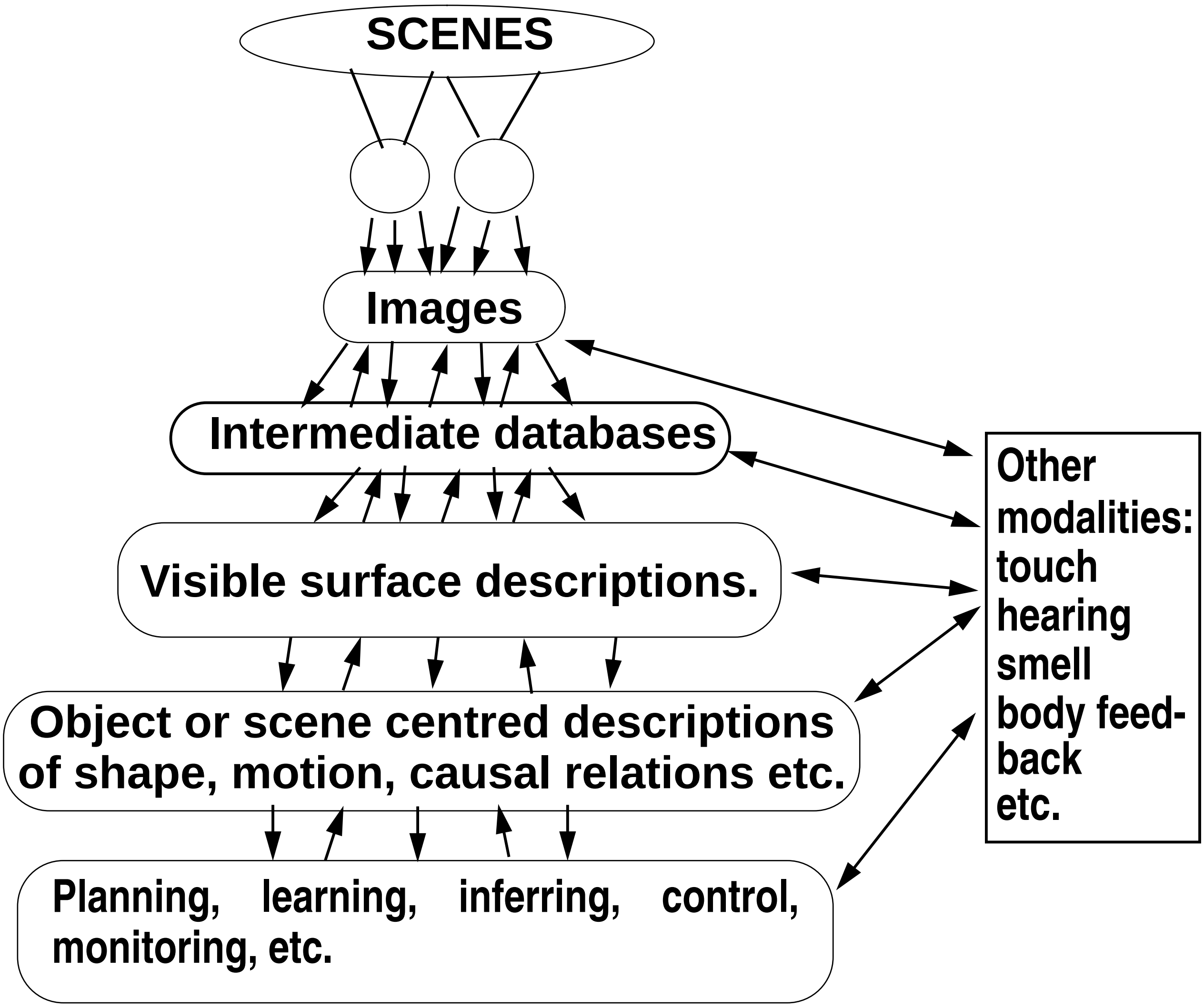
- **Edge-maps, texture-maps, colour maps, intensity maps, etc.**
- **Optical flow**
- **Texture**
- **Histograms of various sorts (Hough transforms)**
- **Databases of edges, lines, regions, binocular disparities, specularity, colour, etc.**
- **Shape from:**
  - intensity and colour variation
  - optical flow
  - texture
  - stereo (binocular disparities)
  - edge contour information
- **Groupings into larger structures**
- **Interpretations in terms of 3-D shape and motion**
- **Construction of relationships:**
  - spatial (inside, next to, touching...)
  - causal (pushing, pulling, pressing, twisting)
  - functional (holding up, keeping shut)
  - intentional (walking towards, picking up, etc.)
- **etc.**

IT'S NOT JUST A RECOGNITION OR LABELLING PROCESS: CREATION AND MAPPING OF STRUCTURES IS ALSO INVOLVED

Perception does not merely label things. There's also explaining ("that's how the clock works"), controlling (e.g. actions in assembling a clock), and many inner reflexes.

PERCEPTUAL CAPABILITIES CHANGE THROUGH LEARNING.

A PARTIAL VIEW OF A VISUAL ARCHITECTURE

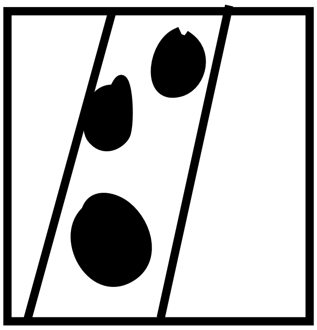


- The internal information structures depend not only on the nature of the environment (E1, E2, etc.) but also on the agent’s needs, purposes, etc. (the Di) and conceptual apparatus. Two organisms, or even two people, can look at the same scene and see different things. Many representational problems are still unsolved.
- Clues to human information structures come from analysing examples in great detail.

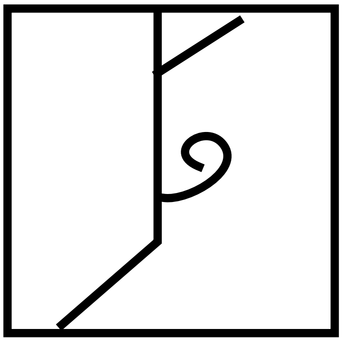
HOW DOES HUMAN VISION WORK?

- **We get some clues by reflecting on what we can see**

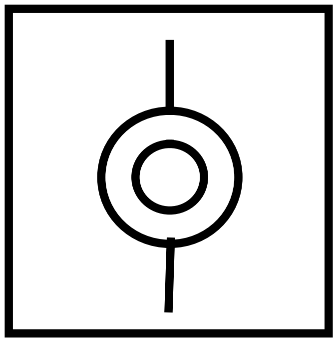
Some “droodles” are radically ambiguous without “top down” hints.



Giraffe walking past the window



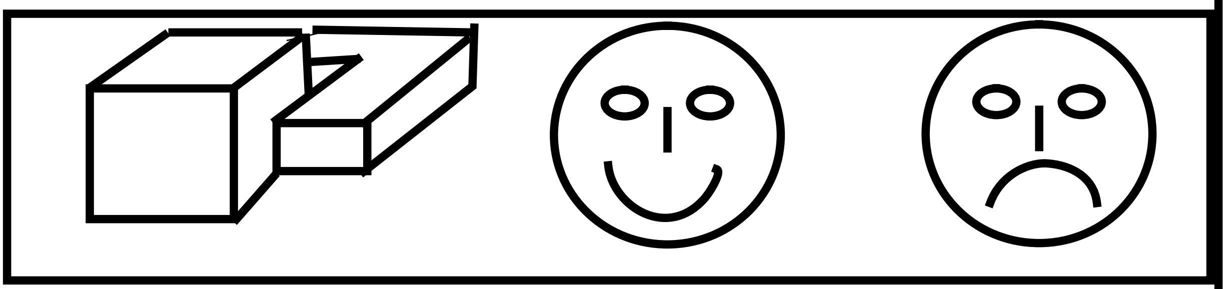
Soldier with gun taking his dog for a walk



Mexican riding a bicycle: where are you?

What structures changed in you when you “saw” what was intended in these droodles?

Some pictures require very rich descriptive resources



Sometimes much deeper and more difficult representations are needed for scenes than for the original images.

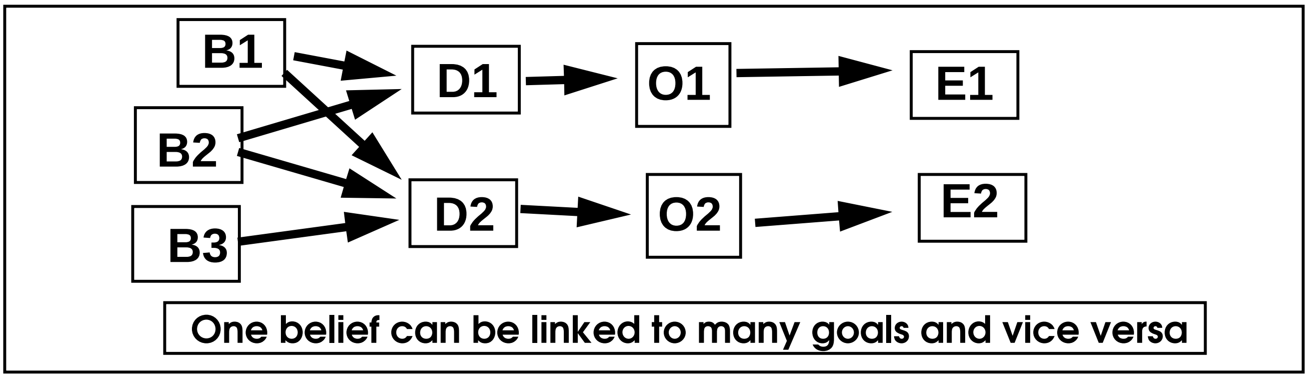
- **The mechanisms and architectures required for an organism that interacts with other intelligent agents must be capable of acquiring and using information about the internal states of others. (What sort of internal state is joy, or pride, or admiration, or sadness, or anger? How is it represented?)**

FURTHER COMPLICATIONS OF DESIGN (2)

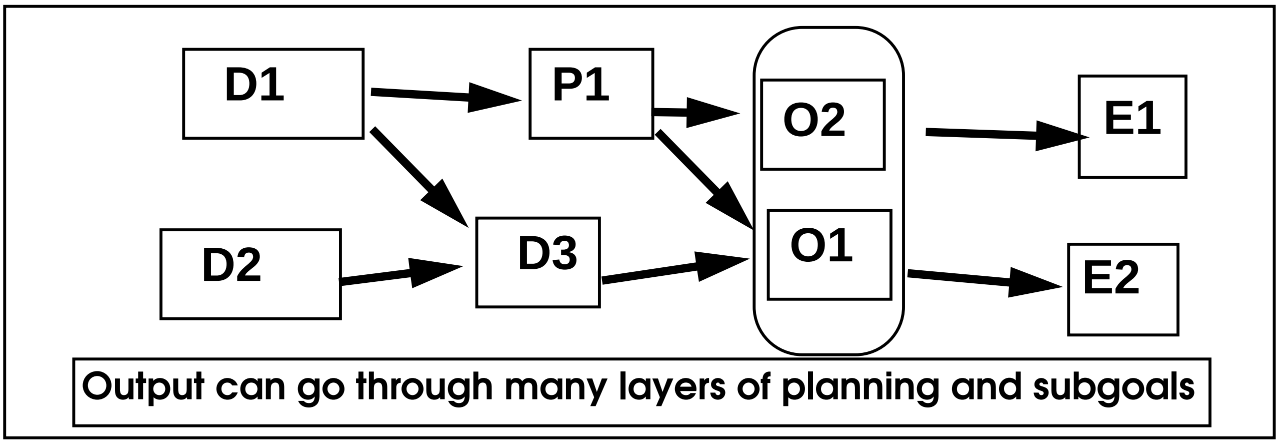
DESIRE-LIKE SUB-STATES (Di)

There are various further ways in which the generation of outputs from desire-like states may be complicated:

- **Information sharing: Particular Di may use several Bi in producing output signals, and particular Bi can be used by many Di (e.g. using many facts in deciding how to achieve one goal; and knowledge about cars can help you drive, and help you avoid being run over)**



- **Causal links between Di and Oi may be indirect, via several layers of causation e.g.**
  (a) going via planning mechanisms, and using different sub-goals to achieve a single goal
  (b) translating high-level to low-level instructions.



-

COMPLICATIONS (2) CONTINUED

- **Some Di change internal states, e.g. other Di and Bi**

  So some control is SELF control.

  E.g. making yourself concentrate on something. In that case some of the Ei are internal. (The mind is part of the environment, for itself) Desires themselves may be produced by deeper or higher level desire-like states (e.g. general attitudes, preferences, etc.) interacting with various Bi to produce new motives. So motivation can involve hierarchies of dispositions. (See next page)
- **Different intermediate Di-controlled sub-states in "output" pathways may use different forms of information storage and transmission. (Compare layers of interpretation of inputs.)**

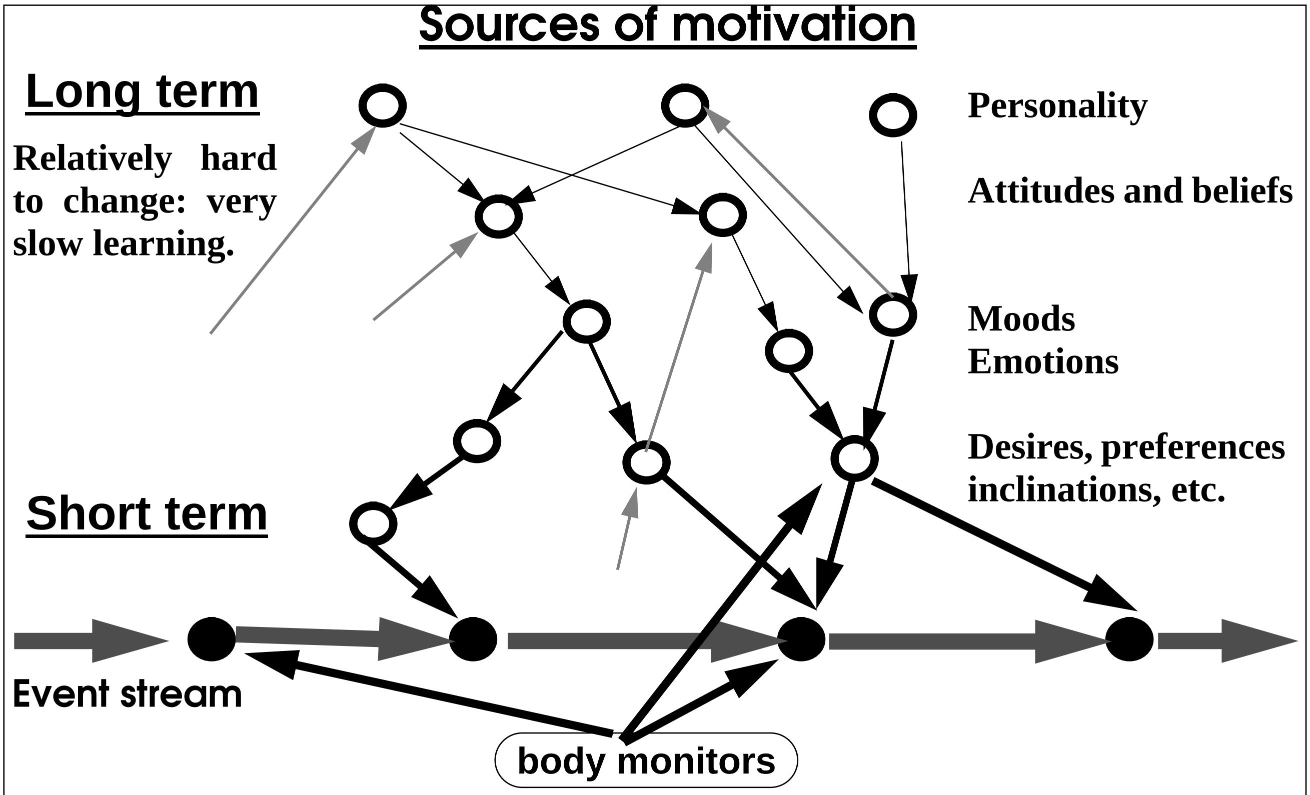
  - E.g. having a thought, shaping a sentence, generating a syntactic form, selecting words, intonation patterns, stress patterns, volume, etc. Compare dancing, sculpting, assembling a clock.
- **The Di need not determine instantaneous output: they may require temporally extended actions. This requires (a) Di states with rich internal struture (e.g. stored plans, with suitable temporary memory mechanisms)**

  (b) "output channels" with considerable sophistication (e.g. program-execution, rule-following, etc.)
- **Some Di are long term dispositions to produce various changes: they don't actually do anything until certain conditions arise. E.g. attitudes like racial prejudice.**
- **Some are "higher level" control states for selecting between conflicting Di (e.g. preferences, principles).**

HIERARCHIES OF DISPOSITIONS

Some dispositions are very long term and hard to change (e.g. personality, attitudes), others more episodic and transient (e.g. desires, beliefs, intentions, moods).

Many are complex, richly-structured, sub-states, e.g. political attitudes. Causal interactions are both context-sensitive (dispositional) and (apparently) probabilistic (propensities, tendencies), not deterministic.



Hierarchies of (dispositional) control

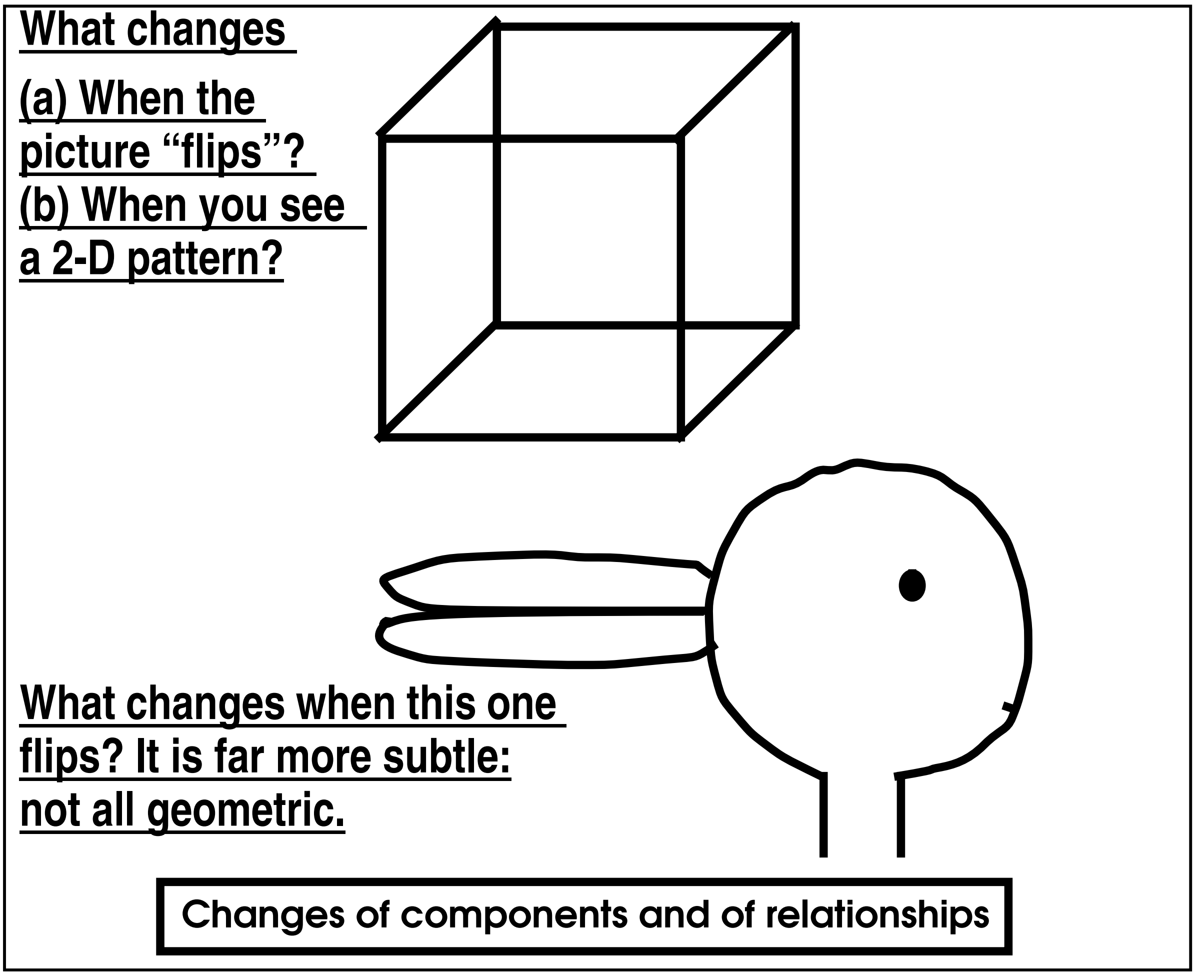
Arrows represent causes of differing strengths, differing time-scales (learning), some deterministic some probabilistic

Engineers know about control hierarchies, but we need richer mechanisms than parameter adjustment. Much change is structural not quantitative (e.g. finding a new plan). Also the “attractor” notion can’t cope with multiple, independent, coexisting dispositions some temporarily suppressed.

FURTHER COMPLICATIONS OF DESIGN (3)

KINDS OF VARIATION

- **Different mechanisms (or parts of one mechanism) provide different kinds of variation. A temperature sensor requires only linear (continuous?) variation. A house-perceiver needs structural variation.**



Kind of variability needed in Bi and Di states depends on both the environment (e.g. does it contain things with different structures?) and the requirements and abilities of the agent. Compare the needs of a fly and of a person. Do flies need to see structures (e.g. for mating)? Do they deliberately create or modify structures? Rivers don’t.

COMPLICATIONS (3) CONTINUED:

The kinds of variability of individual substates (Bi, Di) may be far more sophisticated than in the thermostat, e.g.

- **Multidimensional variation (e.g. sub-states that can be represented as a vector of N independently changeable measures: velocity, position, rotation, etc.)**
- **Structural rather than quantitative variation (e.g. construction of sentence-like, or parse-tree like information stores) requires mechanisms capable of creating and changing structures.**
- **EXAMPLE:**
  - "They make painting machines" vs
  - "They were painting machines" (two readings)
- **Causation between substates includes not only quantities like force, current, torque, but also transmission of structured messages, e.g.**
  - in motive creation,
  - in higher levels of perceptual interpretation,
  - in plan execution.
- **The control architecture itself may need to change as a result of learning. E.g. number and variety of Bi and Di (and other types of control sub-states) change over time, and new causal linkages develop:**
  - A child eventually learns not to let the latest powerful motive dominate. What architectural changes enable the developing child to compare different motives, assess short and long term benefits?
- **Some of the structures, and structural changes occur only in high level virtual machines, e.g. in abstract states of computers or "recurrent" neural nets.**

IMPLICATIONS FOR HUMAN SCIENCES, ENGINEERING, MATHEMATICS EVOLUTIONARY BIOLOGY.

- **Hierarchies of Di, or higher-level Di-producing sub-states implies that (unlike the thermostat) goals don't have to come from outside, or from the "designer" if there is one. They can be produced by a complex system as a result of rich individual development. They are the system's own goals, motives, desires, etc.**
- **Large numbers of active internal causal pathways make the whole system inherently unstable: internal states are constantly in flux, even without external 'stimuli'. MOST "BEHAVIOUR" IS THEN INTERNAL (including changes within virtual machines).**
- **These complications reduce the correspondence between internal Bi and Di states and external states (Ei) and behaviour. Feedback paths can be very complicated and causation can go via multiple routes. So inferring inner states from behaviour is nearly impossible. (Alas poor psychology!)**
- **Time-sharing input and output channels between different Ei and internal states requires various kinds of memory: long term and short term, and different degrees of abstraction. Scheduling is needed: deciding which channel to use when, for what purpose.**
- **Different kinds of attention can be explained in terms of switching patterns of activity: changing what's analysed, or how, and changing what's done, or how. (Design considerations, including learning requirements, may explain limits of human multi-processing capabilities)**

IMPLICATIONS CONTINUED

- **Control is not restricted to parameter adjustment: new structures (new goals, plans, object descriptions) may be created. So, e.g. differential equations are insufficient.**
- **New Maths Needed:The kinds of mathematics developed by control engineers do not seem to be capable of handling the sorts of systems described here. (They can't handle computing systems either: that's not a coincidence. Compare why "general systems theory" didn't work)**
- **Locations of different sub-states at different places in the causal networks correspond to different functional roles of sub-states. Increased complexity of architecture implies increased functional differentiation of sub-states: not just belief-like and desire-like states, but many other kinds (imagining, supposing, planning, attitudes, preferences, principles, personality, etc.)**
- **Functional differentiation (architectural change) can occur both in evolution and in individual development.**
- **There are interesting questions about how coherent control of such a system is possible, and why it doesn't go wrong more often. (Compare multiple personalities, emotional disorders, learning disabilities, etc.)**
- **WHEN THINGS DO GO WRONG, YOU CAN'T HOPE TO BE MUCH GOOD AT HELPING (THERAPY, COUNSELLING, TRAINING) WITHOUT KNOWING THE UNDERLYING DESIGN PRINCIPLES. OTHERWISE IT'S A HIT AND MISS AFFAIR.**
  (I.e. craft, not science or engineering. But some "craft" skills are highly effective, even if we don't know why!)

SOME PHILOSOPHICAL IMPLICATIONS

- **Analysing different processes involving internal self-monitoring (Bi produced by internal Ei) and internal control (high level Di producing internal Ei) may one day sort out the mess of conceptual confusions underlying common notions of “consciousness”. This requires evolution of a new conceptual framework for talking about mental states and processes. (Compare early theories about kinds of stuff.)**
- **Systematic analysis of the functional differentiation of substates and the varieties of processes that are possible could produce a revised vocabulary for kinds of mental states and process. Compare: the periodic table led to a revised vocabulary for kinds of stuff.**
- **Layered interpretation processes using different forms of information store could account for “QUALIA” (which some philosophers believe don’t exist, and others believe can’t be explained in terms of mechanisms).**
- **Systems with the control architecture sketched here will have THEIR OWN goals, desires, etc. Nobody else will have produced them.**
- **Issues concerning “freedom of the will” get solved or evaporated by analysing types and degrees of autonomy within systems so designed.**

EMOTIONS AND RELATED STATES

- **Further development of these ideas and the kinds of states that such systems can get into will show us how many words of ordinary language, including e.g. “emotion” and “mood”, relate to emergent properties of control systems, as saltiness emerges when chlorine and sodium combine. Our vocabulary for describing such states will improve with understanding of mechanisms.**

There are many shallow views about emotional states, including the view that they are essentially concerened with experience of physiological processes. If that were true then anaesthetising the body would be a way to remove grief over the death of a loved one.

A much deeper analysis involves emotions as states of an internal control system: with partial loss of control of mental processes. The grieving mother can’t help thinking back about the lost child, and what she might have done to prevent the death, and what would have happened if the child had lived on, etc. etc.

These are control states of a sophisticated information processing system: physiological processes and feedback are only contingently involved in grief, etc.

{A design-based analysis of the sources of human motivation, and their interactions with other states and processes, are the topic of an ongoing research project in collaboration with Glyn Humphreys and three research students, Luc Beaudoin, Edmund Shing and Tim Read. Liz Robinson,studying children’s views of motivation and emotion, joined recently. We’d like to link up with clinical research also. Related research is being done in a few other places.}

WHAT SORT OF UNDERLYING ENGINE IS NEEDED?

All this is neutral as to what mechanisms are used to implement the various kinds of substates and causal linkages.

They might be neural mechanisms or some other kind. As in circuit design, global properties of the architecture are more important than which particular mechanisms are used, when the design is right.

"ARCHITECTURE DOMINATES MECHANISM"

The detailed mechanisms make only marginal differences as long as they support:

- **sufficient structural variability**
- **sufficient architectural richness**
  - **number of independently variable components**
  - **functional differentiation of components**
  - **variety of causal linkages**
- **sufficient speed of operation**

"Virtual" machines in computers seem to have many of these features. They could be implemented in lower level physical or virtual machines.

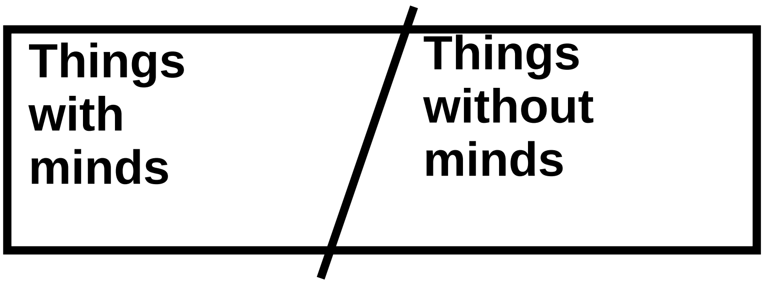
But

WE DON'T KNOW ENOUGH ABOUT REQUIREMENTS, NOR ABOUT AVAILABLE MECHANISMS, TO REALLY SAY YET WHICH INFRASTRUCTURE COULD AND WHICH COULDN'T WORK

E.g. it could turn out that, in our universe, only a mixture of electrical pathways and chemical soup could provide the right combination of fine-grained control, structural variability and global control.

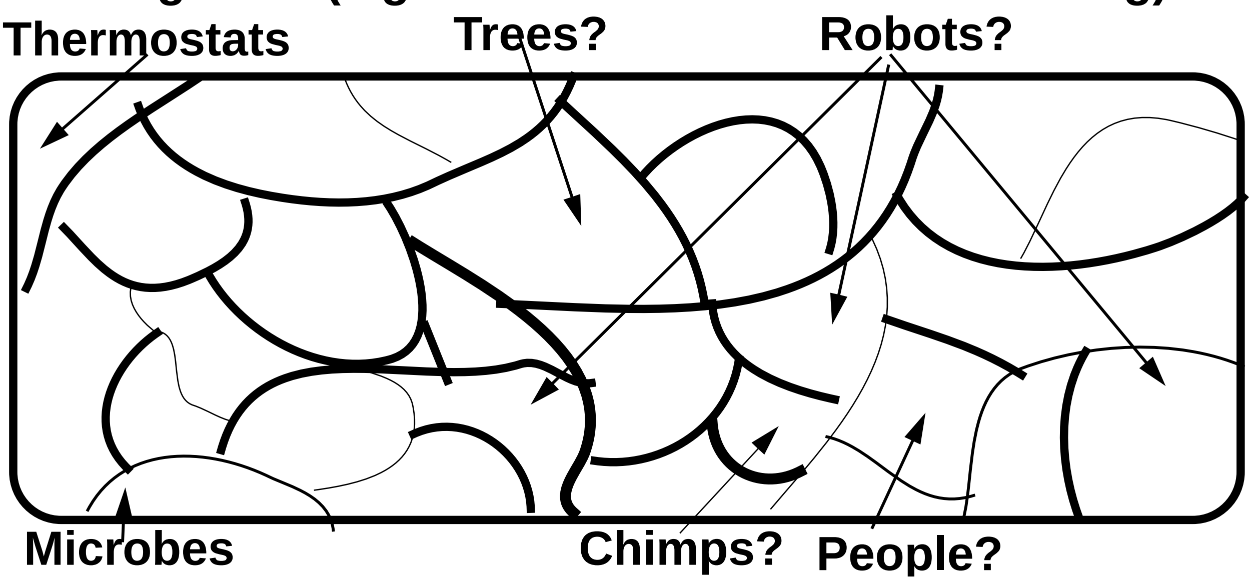
THE SHAPE OF DESIGN SPACE

It's not a dichotomy:



It's not a (smooth or linear) continuum:

There are MANY small but significant discontinuities, and some big ones (e.g. whether there's self-monitoring):



This picture is still too simple: e.g. single-layered. There are still many design options and tradeoffs that we don't yet understand. We need a whole family of new concepts, based on a theory of *design architectures and mechanisms*, to help us understand the relation between structure and capability (form and function).

<u>CONJECTURES</u>

- **The ability to cope with <u>structural</u> variation in information stores was a major evolutionary advance in biological control systems, probably requiring the use of “virtual” machines (in the computer science sense of virtual machine: e.g. the Pascal virtual machine, the Lisp virtual machine).**
- **Other major features of more advanced systems include structural variability of the whole architecture during the development of an individual:**
  - E.g. conceptual development
  - Development of new control systems (e.g. the ability to weigh up long term consequences of desires, or to interleave plans, etc.)
- **Anyone who really understands these issues will come to realise that there’s no “magic” in mind. You may feel you have magical or mystical elements: but so would an intelligent, reflective robot with only partial self-understanding!**
- **There are many potential applications besides the obvious engineering ones: e.g. if you acquire a better understanding of learning, motivation, emotions, etc. in terms of information processing and control systmes, then you can vastly improve procedures in education, psychotherapy, counselling, and teaching psychologists, without having to create intelligent machines to replace us!**
- **We need to explore both individual designs, actual and possible, biological and artificial, and also the shape of design space.**

PROSPECTS AT BIRMINGHAM

Many Schools have potential interests in these topics. Psychology and Computer Science obviously. Also:

- **Mathematics**
- **Philosophy**
- **English**
- **Anthropology and Social Sciences**
- **Earth sciences (expert systems being developed)**
- **Education and Continuing Studies**
- **Electrical and Electronic Engineering**
- **Mechanical and Civil Engineering**
- **Medicine and Dentistry**
- **Others ....**

TWO FORMS OF CONTRIBUTION:

1. **Disciplines that study aspects of intelligence: how it is acquired, learnt, used, represented, etc.**
2. **Disciplines that use knowledge: AI can help to articulate the knowledge, model its use, improve its teaching....**

MANY SEEDS EXIST: CAN WE CULTIVATE THEM, TO PRODUCE AN INTERNATIONAL CENTRE FOR "COMPUTATIONAL EPISTEMOLOGY"?

Cooperative beginnings exist already, e.g.:

- **Vision and image interpretation**
- **The study of motivation and emotions**
- **New degrees in AI and CogSci**
- **Philosophy and AI**

JOIN US!

TYPES OF OPTIMISM ABOUT AI

There are three main types of optimists about the long-term objectives of AI:

1. **Wishful thinkers, or emotionally anti-mystical.** **Those who wish it to succeed (e.g. because it would be shocking, thrilling, a great achievement, etc., or because they are emotionally opposed to mystery and magic). Compare those who want to believe that time travel is possible. Wishful thinking doesn't make them right.**

2. **Ignorant and unimaginative:** **Those who are ignorant and assume without good reason that any processes can be replicated on computers: their concept of the variety of forms of processes is limited by what they already know how to explain or model. People with inadequate imagination may fail to grasp the deep difficulties, and be optimistic for bad reasons. They don't notice the sophistication of children, and squirrels.**

3. **Informed optimists:** **Those with detailed knowledge, who can see that what we've begun to learn is but a beginning and full of promise: we see shapes beckoning in the mists, even though we don't yet see them clearly.**

CONCERNING PROSPECTS FOR AI I HOPE I HAVE TURNED YOU INTO

INFORMED,

CAUTIOUS,

OPTIMISTS!