ASSC7 http://www.cs.memphis.edu/˜assc7/ May/June 2003
Association for the Scientific Study of Consciousness

Architecture-based Philosophy of Mind

What kind of virtual machine is capable of human consciousness?

Aaron Sloman
http://www.cs.bham.ac.uk/˜axs/
School of Computer Science
The University of Birmingham

These slides will be available online at
http://www.cs.bham.ac.uk/research/cogaff/talks/#talk25

See also: the Cognition and Affect Web Site
http://www.cs.bham.ac.uk/research/cogaff/

This is a revised version of a talk presented at ECAP03
http://www.cs.bham.ac.uk/research/cogaff/talks/#talk23
in March 2003 and at The University of Notre Dame in April 2003.

(These slides are still in draft form. Work in progress. Criticisms welcome.)

THEMES

1. Do we know what we mean by “consciousness”?
2. What are information-processing machines?
3. What are virtual machines?
4. How do virtual machines relate to physical machines?
5. How can events in virtual machines be causes, especially of physical events?
6. Atomic State Functionalism *vs* Virtual Machine Functionalism
   i.e. AFM *vs* VMF.
   Property supervenience *vs* mechanism supervenience.
7. Architectures for virtual machines in biological organisms.
8. Consciousness and other aspects of mind as biological phenomena: originally products of evolution, though artificial versions are possible too.
9. Different (virtual machine) information-processing architectures support different varieties of consciousness (and different varieties of motivation, perception, learning, emotion, ...).
10. Some support qualia: the subjects of certain sorts of predication.

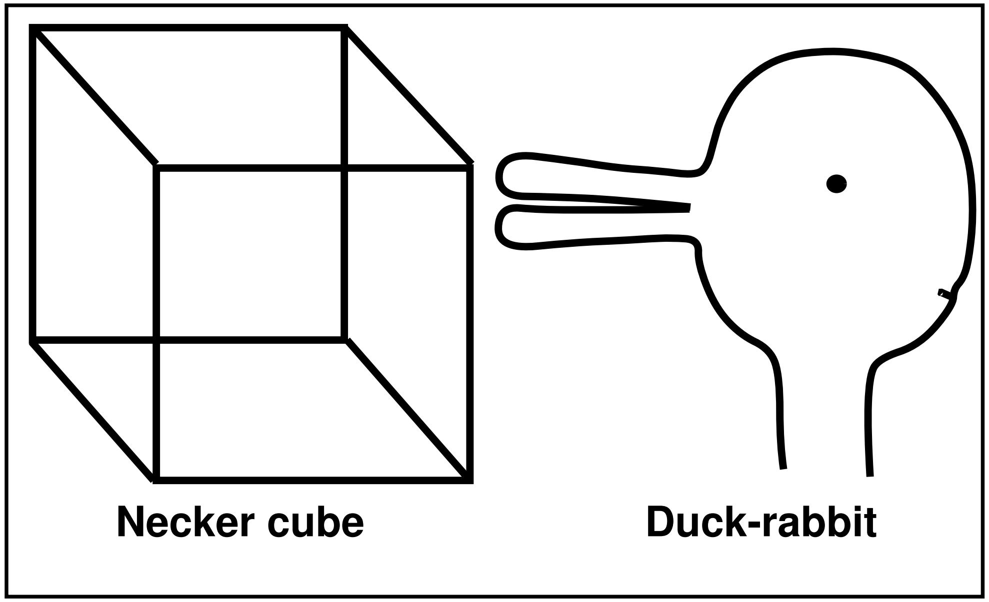
The importance of virtual machines

Many, and probably all, researchers in psychology, brain science, ethology, refer to states, events, processes and entities in virtual machines when they talk about experiences, decisions, intentions, thoughts, learning, feelings, emotions. The concepts used are not definable in the language of the physical sciences but they refer to real phenomena which are are implemented or realised in the physical world. By having a clearer view of what virtual machines are, what they can do, and under what conditions they exist, scientists may come up with better, more complete, more powerful, explanatory theories. This requires adopting the “design stance”.

By clarifying the nature of virtual machines, their relationships to phenomena studied by the physical sciences, and especially their causal powers, we can clarify many old philosophical puzzles and explain why they arise naturally in intelligent, reflective, systems with human-like virtual machines.

This is (in part) an exercise in ontology-building

You need an ontology to see, to experience, to want, to choose, to learn.



Experiencing the Necker flip requires only a spatial/geometric ontology involving 2-D and 3-D structures and relationships.

Experiencing the duck-rabbit flip requires a far more subtle ontology involving categories of animals and their functional parts and their relationships to the environment.

What are the "neural correlates" for the ontology used in experiencing something as FACING TO THE LEFT ?

Does "facing" in an animal involve having the capacity to percieve, to act, etc.?

Ontologies
in scientists
and in
their objects of study.

The point being made is multi-faceted:

- Anything that perceives, observes, processes information, must use an ontology (whether it is appropriate for the task or not).
- So things that are conscious use an ontology which both makes possible and limits the sorts of things of which they can be conscious.
- Likewise scientists studying anything use an ontology and if they use a poor ontology they may ask the wrong questions and come up with poor theories.
- This applies as much to scientists studying consciousness as to all others.

I suggest that much of the scientific study of consciousness makes use of an impoverished ontology, both for describing what needs to be explained and for constructing explanatory theories.

IF THERE ARE NEURAL CORRELATES OF CONSCIOUSNESS

THEY MUST INCLUDE NEURAL CORRELATES FOR THE ONTOLOGIES USED

How is an ontology implemented in a brain?

Some ontologies are relatively simple and can be mapped into two or three (or more) dimensional structures:

E.g.

- An ontology for characterising edge-features in a visual field in terms of location, orientation, contrast, width, etc. could be expressed as a high-dimensional array, where each location represents a vector of values – such things have been found in the visual cortex.
- Additional mechanisms would be required to interpret those vectors, i.e. provide semantics for them.
- In brains, “early” visual regions seem to do this for the simplest, lowest-level visual features.
- A clever interpreting mechanism could even make it scale-invariant.
- Doing it for collections of relationships between low level features such as those suggesting a rabbit silhouette or a duck silhouette is a totally different matter, involving combinatorially explosive search spaces, and as yet unknown solutions to problems of representation.
- Perhaps a generative mechanism could overcome this.

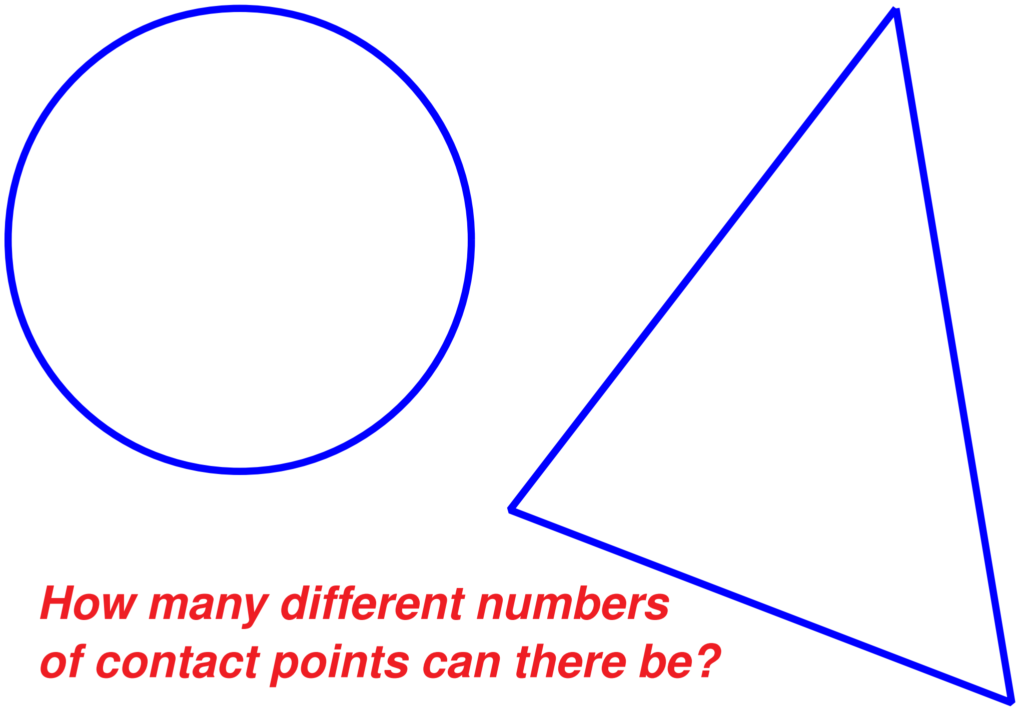
Example: Visual reasoning in humans

Think about this problem:

No points are common to the triangle and the circle. Suppose the triangle changes its size and shape and moves around in the surface, and circle also changes size and location. They could come into contact.

If a vertex touches the circle, or one side becomes a tangent to the circle, there will be one point common to both figures.

If one vertex moves into the circle and the rest of the triangle is outside the circle how many points are common to the circle and the triangle?



How many different numbers of contact points can there be?

How do you answer the question in red?

What forms of representation do you use?

What mechanisms operate on them?

Does the answer change if the circle turns into an ellipse?

What goes on when the circle/triangle question is answered

Most people claim to do it using visualisation, in their heads.

This requires the ability to experience:

- **empty space as containing possible paths of motion,**
- **fixed objects as things that can move, rotate and change their shape,**
- **curves and lines as things that can intersect and touch.**

As well as the ability:

- **to cause various imagined motions**
- **detect and count the resulting contact points**
- **check that all possibilities have been covered.**

The ability to answer the questions requires use of an ontology that includes not only static structures (line, circle, vertex, lines crossing, lines touching) but also kinds of motions on a plane, and the effects that can be caused by such motions.

The ability to check that all possibilities have been covered requires an ontology including classes of phenomena with different features and the ability to check that examples considered suffice to represent every possible class of that sort (e.g. every possible class of configurations giving a different number of contact points). Can any other animal do this?

This is closely related to the perception of affordances.

How is visual problem solving capability implemented in brains?

As far as I know, current methods of investigating such things can at most show which brain regions are active when visualisation is used to solve a problem. Showing where something happens does not explain what is done or how it is done.

Can any other animal solve geometric problems by visualising transformations?

How do nest-building birds (e.g. magpies) decide how to insert the next twig?

Does Betty, the hook-making crow, understand the possibility of bending a straight wire to make a hook before she does it, and if so how does she grasp this possibility? See the video here: http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm

For further discussion, including the problem of understanding how Mr Bean can remove his underpants without removing his trousers, see A. Sloman, (2001) Diagrams in the mind, in *Diagrammatic Representation and Reasoning,* Eds. M. Anderson, B. Meyer & P. Olivier, Springer-Verlag, Berlin, also on CogAff web site.

The paper claims that some of the reasoning switches between metrical and topological relationships. I don't think anyone knows either how these capabilities are implemented in brains or how to give them to machines. Part of the problem is to characterise the capabilities accurately.

Different organisms need different ontologies. Different scientists need different ontologies.

What sort of ontology is needed for a scientific study of consciousness?

An ontology for scientists studying mentality must include, among other things:

- Forms of representation
- Information (semantic, not Shannon-Weaver)
- Information-relations (e.g. contradiction, consistency, entailment, equivalence ...)
- Information-processing mechanisms
- Architectures
- Ontologies
- Control
- Virtual machine
- .... etc ....

See unpublished paper by Sloman and Chrisley on “ontological blindness” in biologists and engineers (e.g. biologically inspired roboticists), on the Cognition and Affect web site.

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

A recent biological development

The evolution of organisms able to discover qualia

We need to understand the virtual machine architecture within which processes occur that involve

attending to and categorising aspects of processes that occur within the system.

It will need (among other things)

- appropriate mechanisms for inspecting some of its own states
- an ontology for categorising them
- a form of representation for expressing the information

We also need to explain why and how such an architecture could evolve: it might have started as a special case of a general biological phenomenon:

Evolved mechanisms often turn out to have effects other than those for which they were selected.

Ontologies used by organisms and by scientists

Just as

- A biologist need not use the same ontology as that used by organisms whose information-processing capabilities are explained by the biologist's theory (e.g. the biologist cannot express in English what the dancing bee communicates to other bees, though it can be described in English.)

So also:

- A scientist fully explaining phenomena of consciousness need not use the same ontology as is used by any particular individual (or species) whose capabilities to reflect on their own consciousness are explained.

And a theory of how ontologies and concepts work will explain why.

E.g. a self-describing machine may use "causally-indexical" concepts (John Campbell's term) to characterise some of its internal virtual machine states.

See:

A. Sloman and R.L. Chrisley (2003) Virtual machines and consciousness, *Journal of Consciousness Studies* 10, 4-5.

NOTE:

Thomas Nagel came close to saying something like this at the end of "What is it like to be a bat?"

We need to go beyond Dennett's "Heterophenomenology" and use the design stance:

understanding architectural possibilities, and using architecture-based concepts, helps us ask sharper questions about what the phenomenology might be.

(See CogAff papers on architecture-based concepts for more examples.)

Caveat

NB:

A scientist producing such an explanation will have to use an ontology appropriate to describing and explaining the functioning of sophisticated information-processing systems.

This ontology is not yet fully articulated: evolution is way ahead of us as it is way ahead of mechanical engineers...

So we have much work to do.

In particular we must learn to tell when we are using confused ontologies and therefore asking pseudo-questions.

Let’s vote!

- **Is a fish conscious?**

Let's vote!

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?

Let's vote!

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?
- Is a new-born baby conscious (when not asleep) ?

Let’s vote!

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?
- Is a new born baby conscious (when not asleep) ?
- Are you conscious when you are dreaming?

Let's vote!

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?
- Is a new born baby conscious (when not asleep) ?
- Are you conscious when you are dreaming?
- Can a five month human foetus be conscious?

Let’s vote!

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?
- Is a new born baby conscious (when not asleep) ?
- Are you conscious when you are dreaming?
- Can a five month human foetus be conscious?
- Is a soccer-playing robot conscious?

Let's vote!

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?
- Is a new born baby conscious (when not asleep) ?
- Are you conscious when you are dreaming?
- Can a five month human foetus be conscious?
- Is a soccer-playing robot conscious?
- Could the robot be conscious of the opportunity to shoot?

Let's vote!

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?
- Is a new born baby conscious (when not asleep) ?
- Are you conscious when you are dreaming?
- Can a five month human foetus be conscious?
- Is a soccer-playing robot conscious?
- Could the robot be conscious of the opportunity to shoot?
- Is the file-protection system in an operating system conscious of attempts to violate access permissions?

Let's vote!

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?
- Is a new born baby conscious (when not asleep) ?
- Are you conscious when you are dreaming?
- Can a five month human foetus be conscious?
- Is a soccer-playing robot conscious?
- Could the robot be conscious of the opportunity to shoot?
- Is the file-protection system in an operating system conscious of attempts to violate access permission?
- Can events in a virtual machine have causal powers?

  We'll return to that later.

Do we know what we mean by "consciousness"?

Many philosophers discuss consciousness as if there were one thing referred to by the noun 'consciousness' and anything either has it or does not have it.

On that view it makes sense to ask questions like

- **"When did IT evolve?"**
- **"Which animals have IT?"**
- **"Which brain mechanisms produce IT?"**
- **"At what stage does a foetus have IT?", etc.**

Do we know what we mean by "consciousness"?

Many philosophers discuss consciousness as if there were one thing referred to by the noun 'consciousness' and anything either has it or does not have it.

On that view it makes sense to ask questions like

- "When did IT evolve?"
- "Which animals have IT?"
- "Which brain mechanisms produce IT?"
- "At what stage does a foetus have IT?", etc.

Problem:

people who share the assumption that they know what they mean by "consciousness" often disagree, not only on

the answers to such questions,

but also on

what sort of evidence could be relevant to answering them.

That’s evidence for deep muddle

Those disagreements suggest that the concept, as used by such philosophers, and also scientists who join in philosophical debates, is full of muddle and confusion – even if there’s nothing wrong with its use by non-professionals to ask and answer questions like:

- **“Is he still unconscious?”**
- **“When did he regain consciousness?”**
- **“Were you conscious that people were looking at you?”**

  etc.

(Non-professionals know how to answer those questions, in most normal contexts.)

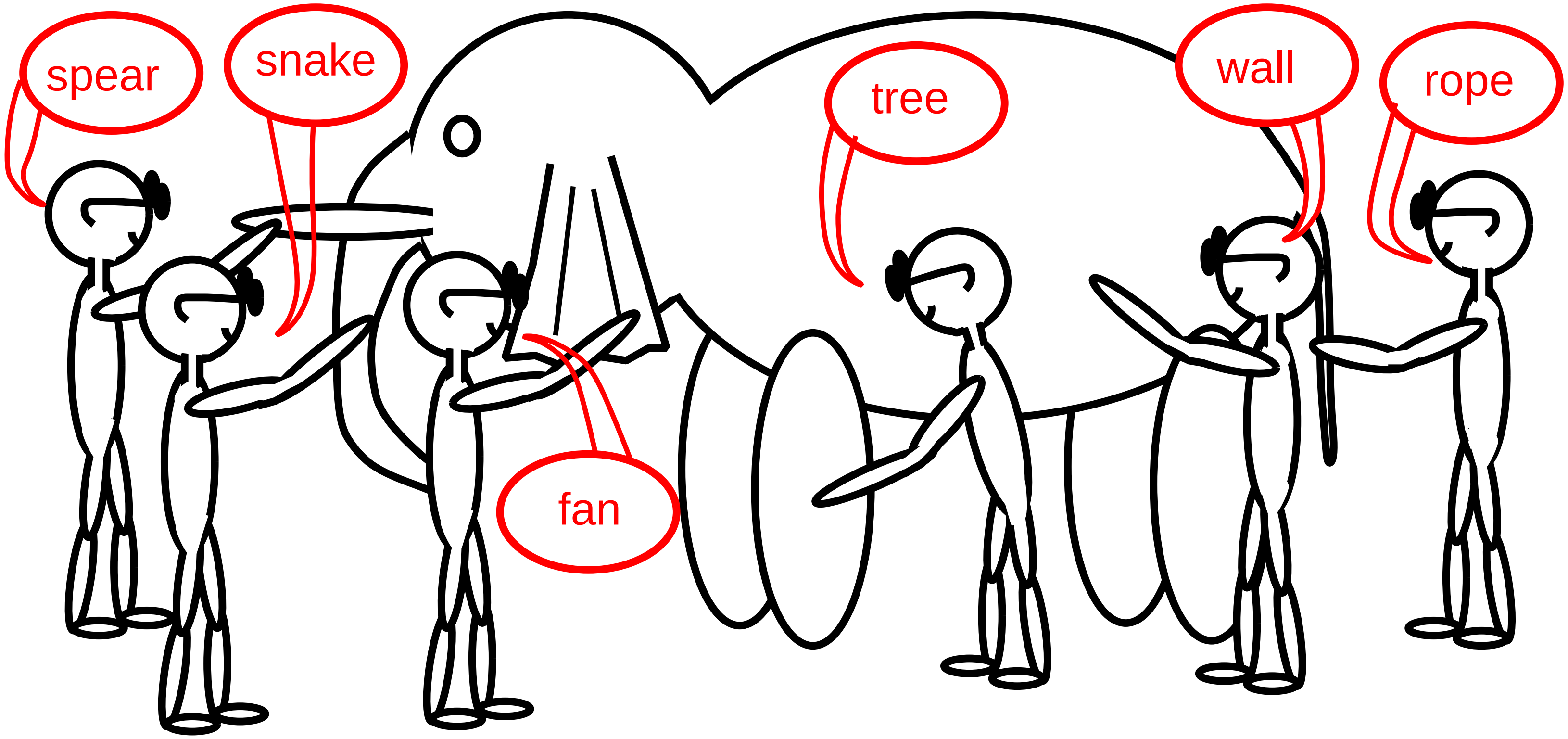
Some professionals studying “consciousness” assume that the disagreements arise because consciousness is just a matter of degree.

There’s another alternative:

> **Human minds include a large and diverse collection of capabilities, and different theorists unwittingly focus on different subsets of those capabilities.**

What capabilities?

Is consciousness an elephant?



See: “The Parable of the Blind Men and the Elephant”
by John Godfrey Saxe (1816-1887)

http://www.wvu.edu/˜lawfac/jelkins/lp-2001/saxe.html

Different theorists focus on different subsets of a very complex and ill-understood reality: the differences are not matters of degree.

Note on next few slides

The next few slides show a rather complex diagram which loosely represents some ideas regarding the information-processing architecture of a typical adult human.

It indicates a high level subdivision of a vast collection of different capabilities.

This architecture is called H-Cogaff, as it is a special case of the generic CogAff architecture-schema which represents a very wide range of possible architectures.

Both CogAff and H-Cogaff are discussed in detail in recent papers in the Cognition and Affect project directory

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

and in slide presentations in this directory:

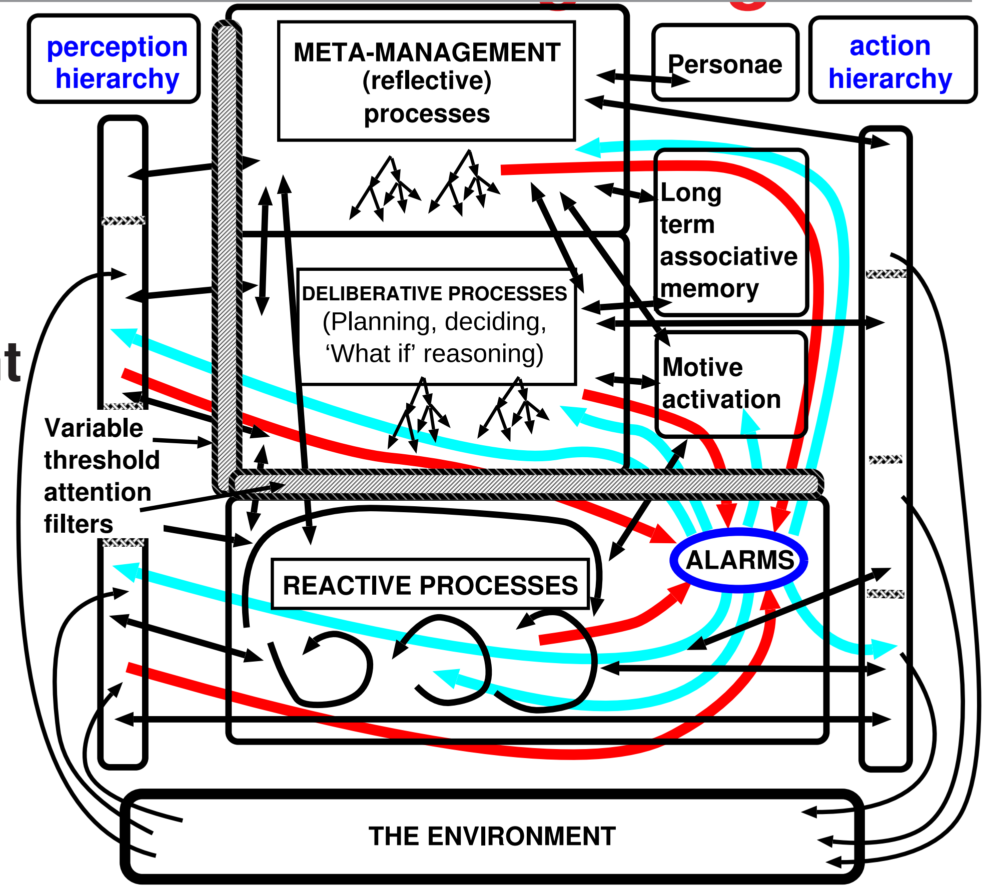
http://www.cs.bham.ac.uk/research/cogaff/talks/

The points made using the architecture do not depend on the details of the H-Cogaff architecture so much as on the fact that it includes

several different subsystems that run concurrently doing different sorts of tasks, though often with some overlap, and often using the same sensory input transformed and processed in different ways.

Some subsystems work much faster than others, when doing the same high level task, but do so with less flexibility, generality, and control (something like the difference between compiled and interpreted versions of the same program running on a computer).
Probably these faster but less flexible sub-systems use evolutionarily old reactive mechanisms.

What capabilities are we investigating?



A large collection of
- **reactive**
- **deliberative**
- **meta-management**

capabilities, that coexist and operate concurrently in different parts of a complex information-processing architecture.

The different capabilities
- **evolved at different times**
- **develop at different times in individuals**
- **manipulate different kinds of information, at different speeds,**
- **use different ontologies and different forms of representation**
- **overlap in what they can do**
- **support different kinds of consciousness, different kinds of motivation, different kinds of learning, different kinds of emotions.**

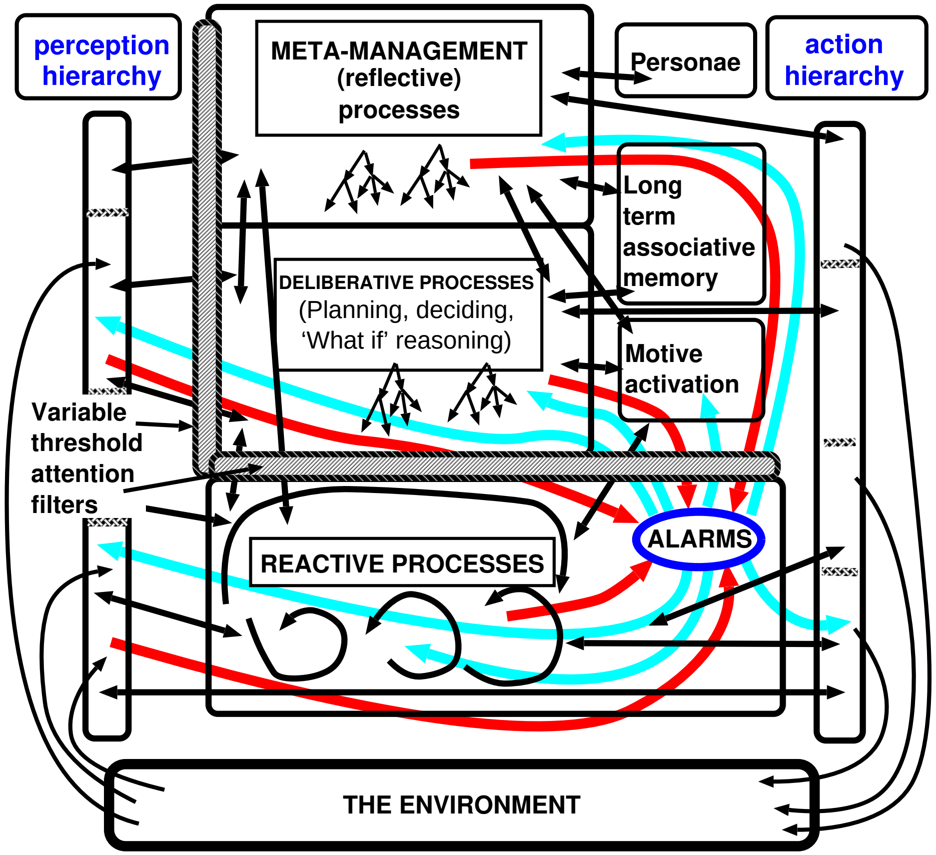
Different theorists unwittingly study different parts of the system.

Whose capabilities?

Theorists (scientists, philosophers) often ask the wrong questions

Don't ask

- What can the subject see?
- Can the subject recognize X?
- When does the subject detect X?
- How does the subject decide to do X?
- Which bit of the brain detects X?



If there are different sub-systems performing partly similar tasks concurrently, there may be different answers for different sub-systems which do X in subtly different ways, using different mechanisms.

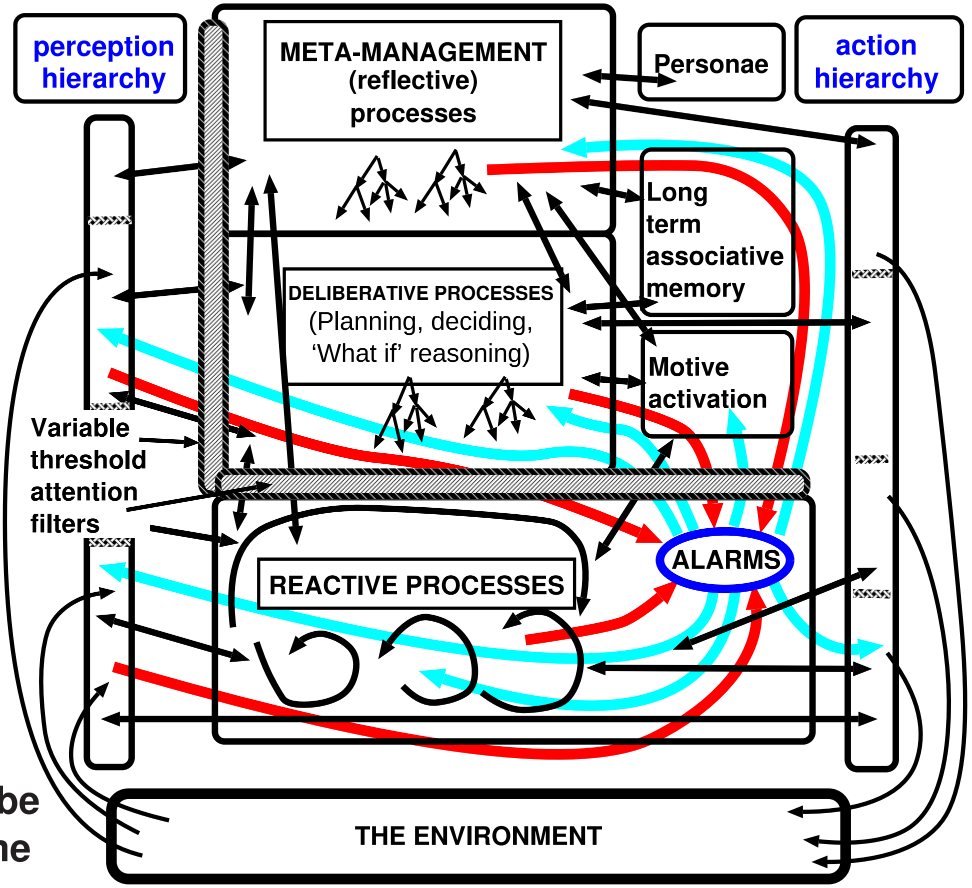
E.g. the same task done by a novice and by a highly trained expert may use different sub-systems, e.g. reading text, sight-reading music, perception of a tennis server, seeing what to grasp while leaping through trees, deciding what to do in each case, etc. The novice has only slow, generic, deliberative mechanisms, while the expert uses highly trained specialised reactive skills.

Capabilities used for what purpose?

Results of processing may be used for different purposes in different sub-systems (Sloman 1989 *Journal of Experimental and Theoretical AI)*

- One bit of the system may use information about perceived optical flow solely for posture control.
- another bit may be able to use the information to control grasping action in a tight feedback loop.
- another bit may be able to use the information in ballistic actions.
- Another bit may be able to use information to answer questions about what is happening in the environment.
- Another bit may be able to use information to answer questions about what happened within the perceiver (e.g. what was sensed).
- All of those capabilities may be extendable through training.

(E.g. linguists learning to attend to and describe features of phonemes that most people in some sense hear but cannot describe.)



Note that the different capabilities may use different ontologies and different forms of representation – though the original inputs are the same. So perceiving X is not one thing: but different things in different sub-systems.

Used when?

Timings may be different for events in different parts of the system.

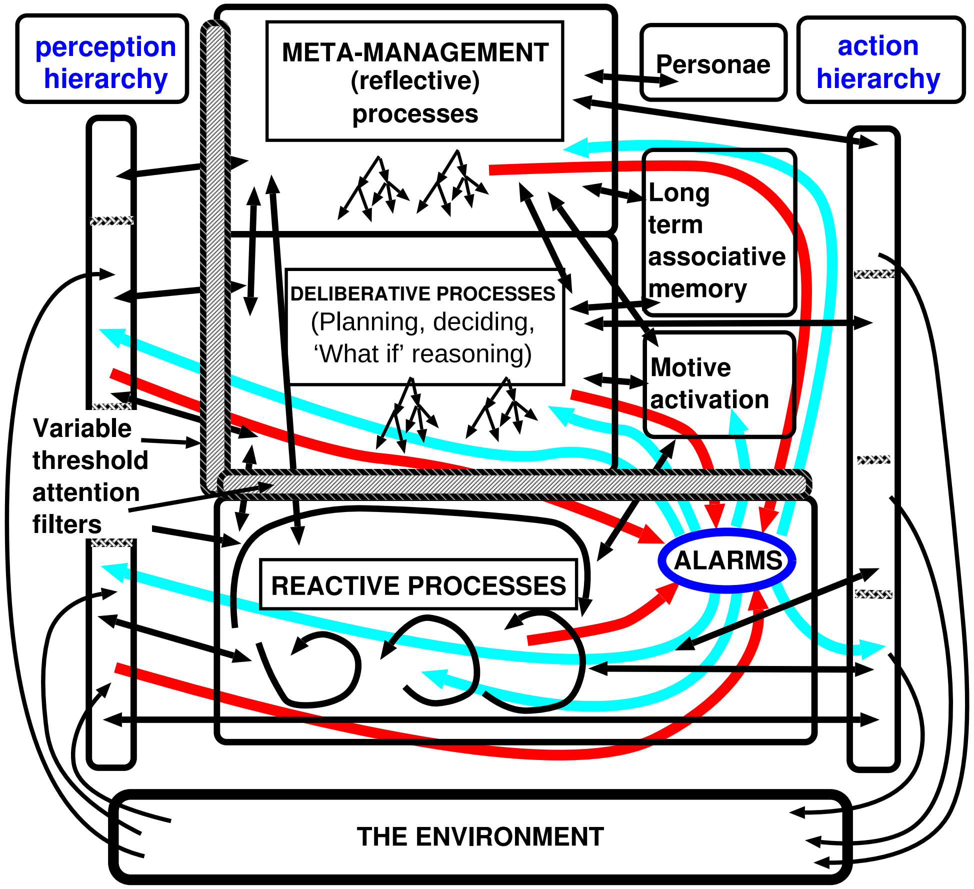
• Asking **exactly when** a particular task is done may be pointless if doing it requires coordination of many sub-processes: *exactly when does a wave reach the sea-shore?*

• It may also not make sense to ask when something occurs if different sub-systems operate on different time-scales

> If a general gives the order to commence battle after some platoons have already started firing in self-defence, when did the battle begin?

• The time at which events happen may depend on whether they are happening naturally in response to detected need, or whether they are in response to verbal instructions, so that a verbally mediated form of control is used, or in response to some external visual or auditory signal.

• The "alarm" mechanism is presumed to be very fast and very inflexible.



All this implies that questions about timing may be either ambiguous or in some cases even meaningless, or the answers may be context-sensitive in subtle ways.

Theatre or Parliament?

- It is commonplace to think of consciousness as some sort of theatre in which various performances occur.
- The considerations presented here suggest that if we are to use such metaphors (perhaps temporarily before coming up with more precise theories) then the metaphor of a parliament may be more appropriate than a theatre metaphor.
- Theatres are too passive - typically they do not decide anything and their contents need not be connected with any external environment.
- Each performance in a theatre is a separate event: there is no progression through life, whereas sessions in a parliament may progressively develop objectives, policies, plans and actions which need to cohere even though they change – a parliament has a memory.
- Theatres operate in too much of a vacuum, whereas parliament has many subcommittees, civil service departments, and information sources, all working in parallel, though in a coordinated fashion, and all serving the needs of a larger whole.

Summary so far

- Research, and understanding, can be seriously hampered if people
  - use an inadequate ontology

    failing to describe accurately what needs to be explained

    or
  - restrict attention to a subset of the phenomena

    like the six blind men.

  when they make observations, formulate hypotheses, ask questions ...

- That's part of the explanation for the problems in the study of consciousness.

But that's only half the problem

A further obstacle to understanding is that most people know about too few modes of explanation of complex processes.

- **Minds are not static entities: processes are going on all the time,**

But that's only half the problem

A further obstacle to understanding is that most people know about too few modes of explanation of complex processes.

- **Minds are not static entities: processes are going on all the time,**
  - some caused by mental events (e.g. decisions),
  - some caused by brain events (e.g. drugs),
  - some caused by perceived physical events,
  - some caused by social events....

But that's only half the problem

A further obstacle to understanding is that most people know about too few modes of explanation of complex processes.

- Minds are not static entities: processes are going on all the time,
  - some caused by mental events (e.g. decisions),
  - some caused by brain events (e.g. drugs),
  - some caused by perceived physical events,
  - some caused by social events....

  - some causing other mental events, e.g. decisions, emotions,
  - some causing physical events, e.g. increased blood flow, grasping, running,
  - some causing social events, e.g. getting married.

But that's only half the problem

A further obstacle to understanding is that most people know about too few modes of explanation of complex processes.

- **Minds are not static entities: processes are going on all the time,**
  - some caused by mental events (e.g. decisions),
  - some caused by brain events (e.g. drugs),
  - some caused by perceived physical events,
  - some caused by social events....

  - some causing other mental events, e.g. decisions, emotions,
  - some causing physical events, e.g. increased blood flow, grasping, running,
  - some causing social events, e.g. getting married.
- **But our understanding of varieties of causation is too limited.**
- **We know about too few kinds of machines.**

But that's only half the problem

A further obstacle to understanding is that most people know about too few modes of explanation of complex processes.

- Minds are not static entities: processes are going on all the time,
  - some caused by mental events (e.g. decisions),
  - some caused by brain events (e.g. drugs),
  - some caused by perceived physical events,
  - some caused by social events....
  - some causing other mental events, e.g. decisions, emotions,
  - some causing physical events, e.g. increased blood flow, grasping, running,
  - some causing social events, e.g. getting married.
- But our understanding of varieties of causation is too limited.
- We know about too few kinds of machines.
- Most people know only about
  - matter-manipulating machines
  - energy-manipulating machines.

But that's only half the problem

A further obstacle to understanding is that most people know about too few modes of explanation of complex processes.

- Minds are not static entities: processes are going on all the time,
  - some caused by mental events (e.g. decisions),
  - some caused by brain events (e.g. drugs),
  - some caused by perceived physical events,
  - some caused by social events....
  - some causing other mental events, e.g. decisions, emotions,
  - some causing physical events, e.g. increased blood flow, grasping, running,
  - some causing social events, e.g. getting married.
- But our understanding of varieties of causation is too limited.
- We know about too few kinds of machines.
- Most people know only about
  - matter-manipulating machines
  - energy-manipulating machines.
- Minds and the phenomena of consciousness do not seem to fit what we know about such machines.

But that’s only half the problem

A further obstacle to understanding is that most people know about too few modes of explanation of complex processes.

- Minds are not static entities: processes are going on all the time,
  - some caused by mental events (e.g. decisions),
  - some caused by brain events (e.g. drugs),
  - some caused by perceived physical events,
  - some caused by social events....
  - some causing other mental events, e.g. decisions, emotions,
  - some causing physical events, e.g. increased blood flow, grasping, running,
  - some causing social events, e.g. getting married.
- But our understanding of varieties of causation is too limited.
- We know about too few kinds of machines.
- Most people know only about
  - matter-manipulating machines
  - energy-manipulating machines.
- Minds and the phenomena of consciousness do not seem to fit what we know about such machines.

We need to understand another class of machines

Another example of finding the right ontology.

Beyond matter-manipulating and energy-manipulating machines

We need to understand a third class of machines:

- information-processing machines,
- especially virtual information-processing machines.

THESE PROVIDE A NEW APPROACH TO THE PROBLEM

- Software engineers have a deep intuitive understanding of the new mode of explanation, but often cannot articulate it.

Beyond matter-manipulating and energy-manipulating machines

We need to understand a third class of machines:

- **information-processing machines,**
- **especially virtual information-processing machines.**

THESE PROVIDE A NEW APPROACH TO THE PROBLEM

- **Software engineers have a deep intuitive understanding of the new mode of explanation, but often cannot articulate it.**
- **Most philosophers know little about it, because of inadequacies in our educational system!**

Beyond matter-manipulating and energy-manipulating machines

We need to understand a third class of machines:

- **information-processing machines,**
- **especially virtual information-processing machines.**

THESE PROVIDE A NEW APPROACH TO THE PROBLEM

- **Software engineers have a deep intuitive understanding of the new mode of explanation, but often cannot articulate it.**
- **Most philosophers know little about it, because of inadequacies in our educational system!**
- **Most people frequently interact with virtual machines, or indirectly depend on them, whether they know it or not:**
  - spelling checkers
  - email programs
  - games software, e.g. a chess virtual machine
  - document formatters
  - spam filters
  - process-schedulers
  - file-system managers with privilege mechanisms
  - control systems for chemical plants or airliners.

Two notions of virtual machine

Some people object to the idea that causal interactions can occur in a virtual machine, or that events in a virtual machine can be caused by or can cause physical events, because they ignore the difference between:

- **a VM which is an abstract mathematical object**
  (e.g. the Prolog VM, the Java VM)
- **a VM that is a running instance of such a mathematical object, controlling events in a physical machine.**

Two notions of virtual machine

Some people object to the idea that causal interactions can occur in a virtual machine, or that events in a virtual machine can be caused by or can cause physical events, because they ignore the difference between a VM which is an abstract mathematical object (e.g. the Prolog VM, the Java VM) and a VM that is a running instance of such a mathematical object, controlling events in a physical machine.

Physical processes:	Running virtual machines:	Mathematical models:
currents	calculations	numbers
voltages	games	sets
state-changes	formatting	grammars
transducer events	proving	proofs
cpu events	parsing	Turing machines
memory events	planning	TM executions

Two notions of virtual machine

Some people object to the idea that causal interactions can occur in a virtual machine, or that events in a virtual machine can be caused by or can cause physical events, because they ignore the difference between a VM which is an abstract mathematical object (e.g. the Prolog VM, the Java VM) and a VM that is a running instance of such a mathematical object, controlling events in a physical machine.

Physical processes:	Running virtual machines:	Mathematical models:
currents	calculations	numbers
voltages	games	sets
state-changes	formatting	grammars
transducer events	proving	proofs
cpu events	parsing	Turing machines
memory events	planning	TM executions

VMs as mathematical objects are much studied in meta-mathematics and theoretical computer science. They are no more causally efficatious than numbers.

The main theorems, e.g. about computability, complexity, etc. are primarily about mathematical entities (and non-mathematical entities with the same structure – but no non-mathematical entity can be proved to have any mathematical properties).

Two kinds of abstractions: three kinds of machines

We've seen that in addition to physical machines we can have two kinds of abstract machines: mathematical models and running virtual machines.

- Physical machines and virtual machines running in physical machines actually DO things:

  a calculation in a VM, or the reformatting of text in a word-processor, or a decision to turn a valve on
  - can cause other things to change in the VM
  - and can also cause physical events and processes – controlling machinery.

  Sometimes they don't do what was intended, and bugs in the virtual machine have to be discovered and eliminated: much of the work of software engineers is like that – there need not be any fault in a physical component in such cases.

Two kinds of abstractions: three kinds of machines

We've seen that in addition to physical machines we can have two kinds of abstract machines: mathematical models and running virtual machines.

- **Physical machines and virtual machines running in physical machines actually DO things:**

  a calculation in a VM, or the reformatting of text in a word-processor, or a decision to turn a valve on
  - **can cause other things to change in the VM**
  - **and can also cause physical events and processes – controlling machinery.**

  Sometimes they don't do what was intended, and bugs in the virtual machine have to be discovered and eliminated: much of the work of software engineers is like that – there need not be any fault in a physical component in such cases.

- **The mathematical machines (e.g. unimplemented TMs) are abstract objects of study, but they no more act on anything in the world than numbers do, though they can help us reason about things that do act on the world, which they model, as equations can, for instance.**

As many have noted, causation in virtual and physical machines involves a kind of causal circularity. Instead of trying to deny it we need to understand it: e.g. if we are to design, build, use, explain and trust virtual machines.

(For experts:) Two sorts of ‘running’ virtual machines

The situation is confusing if we ignore the differences between compiled and interpreted programs on computers.

- If some AI program AIP is running in a computer, as a compiled machine-code program, then it is possible that the compiled program does not go through operations of the sorts specified in the source code, e.g. because an optimising compiler has transformed the program, or because some arcane sequence of bit manipulations happens to produce the required input-output mapping.
- If AIP is stored in something close to its original form (e.g. as a parse tree) and then interpreted, the various portions of the program are causally effective insofar as they determine the behaviour produced by the interpreter: if they are changed then the behaviour changes, which will not happen if source code of a compiled program is changed.

  (Incremental compilers complicate matters, but will not be discussed here.)
- Thus if we say a program written in a batch-compiled language like C++ uses the C++ virtual machine, there is a sense in which the C++ instructions themselves have no effect at run-time, for they are replaced by machine instructions. However there could be data-structures interpreted as rules which do affect the running, e.g. rules for a game interpreted by a C++ program.
- So deciding whether a particular VM is actually running on a machine or whether it is something else that simulates it that is running, can be tricky. It all hangs on which causal interactions exist in the running VM.

Levels (virtual machines) in reality

At all levels there are objects, properties, relations, structures, mechanisms, states, events, processes and CAUSAL INTERACTIONS.

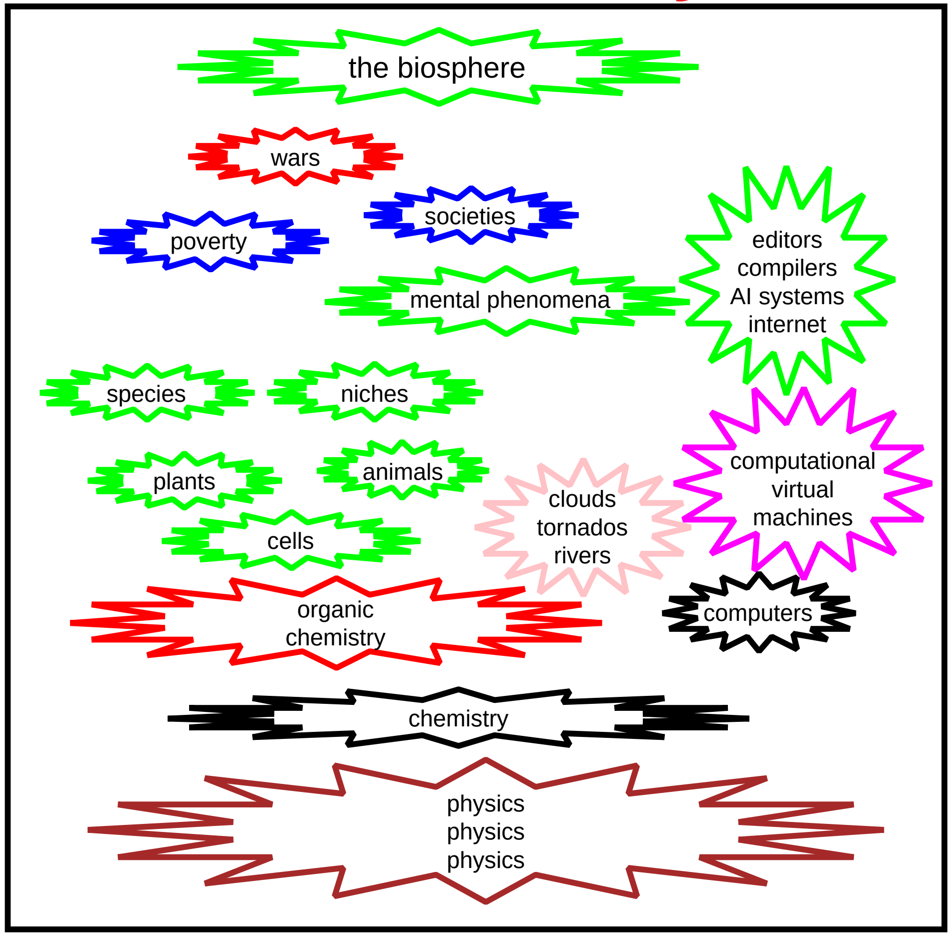
E.g. poverty can cause crime.

But they are all ultimately realised (implemented) in physical systems.

Different disciplines use different approaches (not always good ones).

Nobody knows how many levels of virtual machines physicists will eventually discover. (uncover?)

Our emphasis on virtual machines is just a special case of the general need to describe and explain virtual machines in our world.



See the IJCAI'01 Philosophy of AI tutorial (written with Matthias Scheutz) for more on levels and causation:

http://www.cs.bham.ac.uk/˜axs/ijcai01/

Beyond correlation and causation to implementation/realisation

When a virtual machine (e.g. a spelling checker, operating system, learning mechanism, flight controller) runs on a physical machine, it is a mistake to assume that the relationship between the virtual machine entities and the physical machine components is

- Simply correlation
- Simply causation
- A kind of identity (which normally implies symmetry).

Rather it is a relationship which may be called "implementation" or "realisation" that software engineers understand intuitively because all their work depends on at least a partial understanding of it, but which has never been adequately analysed "at a philosophical level".

The relationship between virtual and physical machines in computers involves setting up a lot of hardware and software components to maintain the truth of a complex collection of counterfactual conditionals (about "what would happen if") which ensure that certain causal connections hold between virtual machine events and physical machine events.

For further discussion on this see Talks 5, 12, 22 here:
http://www.cs.bham.ac.uk/research/cogaff/talks/

Information-processing virtual machines

- A machine is a complex entity with parts that interact causally so as to produce combined effects, either within the parts or externally to the machine.

Information-processing virtual machines

- A machine is a complex entity with parts that interact causally so as to produce combined effects, either within the parts or externally to the machine.
- A virtual machine, or abstract machine, is one whose components are not describable using the language of the physical sciences (physics, chemistry, etc.) but which depends on the existence of a physical machine in order to operate.

Information-processing virtual machines

- A machine is a complex entity with parts that interact causally so as to produce combined effects, either within the parts or externally to the machine.
- A virtual machine, or abstract machine, is one whose components are not describable using the language of the physical sciences (physics, chemistry, etc.) and which depends on the existence of a physical machine in order to operate.

We have only recently begun to understand what virtual machines are.

Some occur naturally in organisms, all of which process information.

These have existed for millions of years.

Information-processing virtual machines

- A machine is a complex entity with parts that interact causally so as to produce combined effects, either within the parts or externally to the machine.
- A virtual machine, or abstract machine, is one whose components are not describable using the language of the physical sciences (physics, chemistry, etc.) and which depends on the existence of a physical machine in order to operate.

We have only recently begun to understand what virtual machines are.

Some occur naturally in organisms, all of which process information.

These have existed for millions of years.

Recently we have begun to learn how to design, implement, analyse, debug and explain artificial virtual machines, e.g.

- Computer operating systems
- Word processors
- Chess playing machines
- Email systems
- Compilers
- Spelling checkers
- Artificial neural nets.

Demonstrations of virtual machines

Teasing two (simple) virtual robots

Braitenberg vehicles

Simulated sheepdog

Simplified Shrdlu

BIOLOGICAL INFORMATION PROCESSING

Biologists are used to thinking of genes as carrying information, and reproduction as transfer of information. Information controls growth from an egg or seed. Organisms acquire and use information in order to survive, reproduce, find shelter, etc.

Most (or all) biological processes,
including perception, learning, choosing, and behaving,
involve acquisition, processing and use of information.

BIOLOGICAL INFORMATION PROCESSING

Biologists are used to thinking of genes as carrying information, and reproduction as transfer of information. Information controls growth from an egg or seed. Organisms acquire and use information in order to survive, reproduce, find shelter, etc.

Most (or all) biological processes,
including perception, learning, choosing, and behaving,
involve acquisition, processing and use of information.

There are different kinds of information, for instance:

- about categories of things (big, heavy, small, red, blue, prey, predator)
- about generalisations (heavy things are harder to pick up)
- about particular things (that thing is heavy)
- about evaluation (X is good, pleasant, etc. Y is bad, unpleasant, etc.)
- about priorities (it is better to X than to Y)
- about what to do (run! fight! freeze! look! attend! decide now!)
- about how to do things (find a tree, jump onto it, climb...)

BIOLOGICAL INFORMATION PROCESSING

Biologists are used to thinking of genes as carrying information, and reproduction as transfer of information. Information controls growth from an egg or seed. Organisms acquire and use information in order to survive, reproduce, find shelter, etc.

Most (or all) biological processes,
including perception, learning, choosing, and behaving,
involve acquisition, processing and use of information.

There are different kinds of information, for instance:

- about categories of things (big, heavy, small, red, blue, prey, predator)
- about generalisations (heavy things are harder to pick up)
- about particular things (that thing is heavy)
- about evaluation (X is good, pleasant, etc. Y is bad, unpleasant, etc.)
- about priorities (it is better to X than to Y)
- about what to do (run! fight! freeze! look! attend! decide now!)
- about how to do things (find a tree, jump onto it, climb...)

Some of these include referential information, some control information, and some both. All presuppose an ontology

We still know only about a small subset of possible types of information, types of encoding, and types of uses of information.

WARNING:

Don't expect all types of information to be expressible in languages WE can understand –

e.g. what a fly sees, or a bee dances!

What is information? What is energy?

The concept of "information" is partly like the concept "energy".

It is hard to define "energy" in a completely general way.

Did Newton understand what energy is?
There are many kinds he did not know about.

We can best think of energy in terms of:

- **the different forms it can take,**
- **the ways in which it can be transformed, stored, transmitted, or used,**
- **the kinds of causes and effects that energy transformations have,**
- **the many different kinds of machines that can manipulate energy**
- **....**

If we understand all that, then we don't need to *define* "energy".

It is a primitive theoretical term – implicitly defined by the processes and relationships that involve it.

We should not use currently known forms of energy to *define* it, since new forms of energy may turn up in future.

Newton knew about energy, but did not know anything about the energy in mass:
Einstein's equation $E = MC^2$ had not been thought of.

Perhaps new forms of energy are yet to be discovered.

Requirements for understanding "information"

Just as understanding what energy is involves knowing many facts about it, likewise knowing what information is involves knowing many facts about it:

- **the different types of information,**
- **the different forms in which they can be expressed,**
- **the different ways information can be acquired, transformed, stored, searched, transmitted or used,**
- **the kinds of causes that produce events involving information,**
- **the kinds of effects information manipulation can have,**
- **the many different kinds of machines that can manipulate information,**
- **the variety of *architectures* into which information processing mechanisms can be combined**

**If we understand all that,
then we don't need to *define* "information"!**

Like "energy", "information", in the sense we use, is an implicitly defined primitive theoretical term.

This is not the Shannnon-Weaver mathematical notion of information, which does not include reference, truth, falsity, contradiction, inference, interpretation....

Information and measurement

One big difference between energy and information: it is very useful to *measure* energy e.g. because it is conserved.

But measuring information (in the sense considered here) is usually less useful – or even meaningless:

> **how much information is on this page?**
>
> **What has more information: a list of Peano's five axioms for arithmetic, or a list of the first billion integers?**

- **I give you information, yet I still have it, unlike energy.**
- **You can derive new information from old, and still have both.**
- **Information varies primarily not in its *amount*, like energy, but in its *structure* and *content*.**
- **Equations do not adequately represent most processes involving manipulation of information.**

Numbers (measurements) do not capture what is most important about information, for behaving systems.

Examples of types of processes involving information

- Acquisition
- Filtering/selecting
- Transforming/interpreting/disambiguating
- Compressing/generalising/abstracting
- Deriving (making inferences, but not only using propositions)
- Storage/Retrieval (many forms: exact, pattern-based, fuzzy)
- Training, adaptation (e.g. modifying weights, inducing rules)
- Mining (for instances, for patterns or rules)
- Constructing (e.g. descriptions of new situations or actions)
- Comparing and describing information (meta-information)
- Reorganising (e.g. formation of new ontologies)
- Testing/interrogating (is X in Y, is A above B, what is the P of Q?)
- Copying/replicating
- Syntactic manipulation of information-bearing structures
- Translating between forms, e.g. propositions, diagrams, weights
- Controlling/triggering/modulating behaviour (internal, external)
- Propagating (e.g. in a semantic net, or neural net)
- Transmitting/communicating .... (and many more)

NOTE: A machine or organism may do some of these things internally, some externally, and some in cooperation with others.
The processes may be discrete or continuous (digital or analog).

Functionalism

The states and processes in a virtual machine have causal powers.

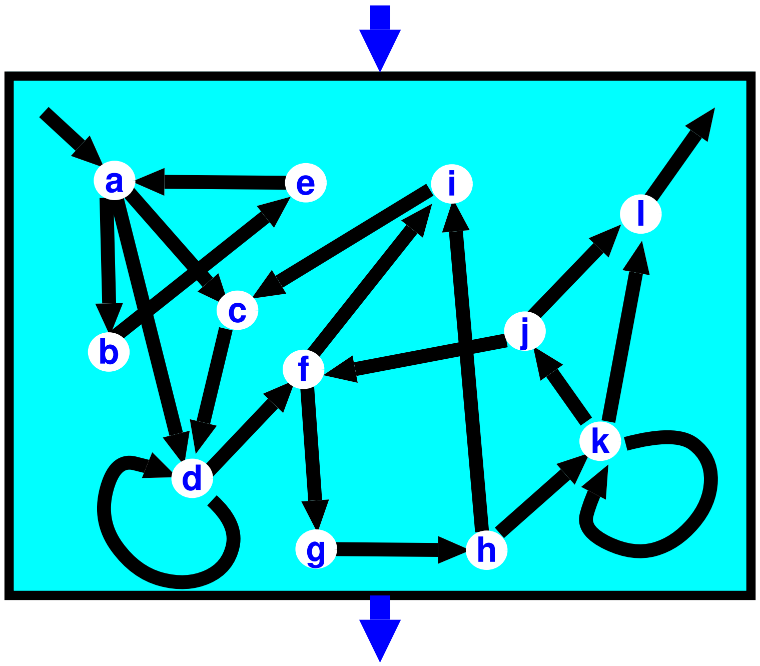
Different such states and different types of mechanisms within the virtual machine can normally be characterised entirely in terms of those causal powers.

But you need to use a good ontology for causal powers:

assuming that they all reduce to input-output relationships is a mistake.

Varieties of functionalism

Some philosophers have attempted to explain what virtual machines are by talking about machines that have states that can enter into causal interactions with inputs, outputs and preceding or succeeding states. The simplest kind is a Finite State Machine (FSM)



This has a collection of possible states (a, b, c, ...) and can receive some sort of input signal and produce an output signal.

Each state is totally defined by its transition rules.

Conventional (atomic state) functionalism

This view assumes that in a physical system there can be only one virtual machine state at a time, and all such states are fully defined by their causal powers, summarised as state-transition rules.

E.g. Block writes:

> "According to functionalism, the nature of a mental state is just like the nature of an automaton state: constituted by its relations to other states and to inputs and outputs. All there is to S1 is that being in it and getting a 1 input results in such and such, etc. According to functionalism, all there is to being in pain is that it disposes you to say 'ouch', wonder whether you are ill, it distracts you, etc."

Human mental states are not like that

Human states like hunger, thirst, puzzlement or anger, do not fit this specification, since these can coexist and start, stop or change independently.

Coexistence of interacting sub-states **is a feature of how people normally view mental states, for instance when they talk about conflicting desires or attitudes, or inferring something new from old beliefs: the old beliefs are stll there.**

Moreover, those states can each have complex internal structures.

- **E.g. wanting to eat an apple includes**
- **having concepts (eat, apple) using them in forming a propositional content,**
- **having that content in mind in a way that tends to cause new thoughts, decisions and behaviours (and whatever else desires do).**

Virtual Machine Functionalism (VMF)

If a mind includes many enduring coexisting, independently varying, causally interacting, states and processes then it is a complex machine with (non-physical) parts that have varying life spans, and which interact causally, e.g. desires, memories, percepts, beliefs, attitudes, pains, etc. (and other things for which there are not ordinary language labels.)

Virtual Machine Functionalism (VMF)

If a mind includes many enduring coexisting, independently varying, causally interacting, states and processes then it is a complex machine with (non-physical) parts that have varying life spans, and which interact causally, e.g. desires, memories, percepts, beliefs, attitudes, pains, etc. (and other things for which there are not ordinary language labels.)

To accommodate this, *virtual machine functionalism* (VMF), is defined to allow

- **multiple,**
- **coexisting,**
- **concurrently active,**
- **constantly changing,**
- **interacting**

mental states.

Virtual Machine Functionalism (VMF)

If a mind includes many enduring coexisting, independently varying, causally interacting, states and processes then it is a complex machine with (non-physical) parts that have varying life spans, and which interact causally, e.g. desires, memories, percepts, beliefs, attitudes, pains, etc. (and other things for which there are not ordinary language labels.)

To accommodate this, *virtual machine functionalism* (VMF), is defined to allow

- **multiple,**
- **coexisting,**
- **concurrently active,**
- **constantly changing,**
- **interacting**

mental states.

Each sub-state S is defined by its causal relationships to other sub-states and, in some cases, its causal relations to the environment. (e.g., if S is influenced by sensors or if it can influence motors or muscles).

Virtual Machine Functionalism (VMF)

If a mind includes many enduring coexisting, independently varying, causally interacting, states and processes then it is a complex machine with (non-physical) parts that have varying life spans, and which interact causally, e.g. desires, memories, percepts, beliefs, attitudes, pains, etc. (and other things for which there are not ordinary language labels.)

To accommodate this, *virtual machine functionalism* (VMF), is defined to allow

- **multiple,**
- **coexisting,**
- **concurrently active,**
- **constantly changing,**
- **interacting**

mental states.

Each sub-state S is defined by its causal relationships to other sub-states and, in some cases, its causal relations to the environment. (e.g., if S is influenced by sensors or if it can influence motors or muscles).

Exactly which states and sub-systems can coexist in humans is an empirical question.

Virtual Machine Functionalism (VMF)

If a mind includes many enduring coexisting, independently varying, causally interacting, states and processes then it is a complex machine with (non-physical) parts that have varying life spans, and which interact causally, e.g. desires, memories, percepts, beliefs, attitudes, pains, etc. (and other things for which there are not ordinary language labels.)

To accommodate this, *virtual machine functionalism* (VMF), is defined to allow

- **multiple,**
- **coexisting,**
- **concurrently active,**
- **constantly changing,**
- **interacting**

mental states.

Each sub-state S is defined by its causal relationships to other sub-states and, in some cases, its causal relations to the environment. (e.g., if S is influenced by sensors or if it can influence motors or muscles).

Exactly which states and sub-systems can coexist in humans is an empirical question.

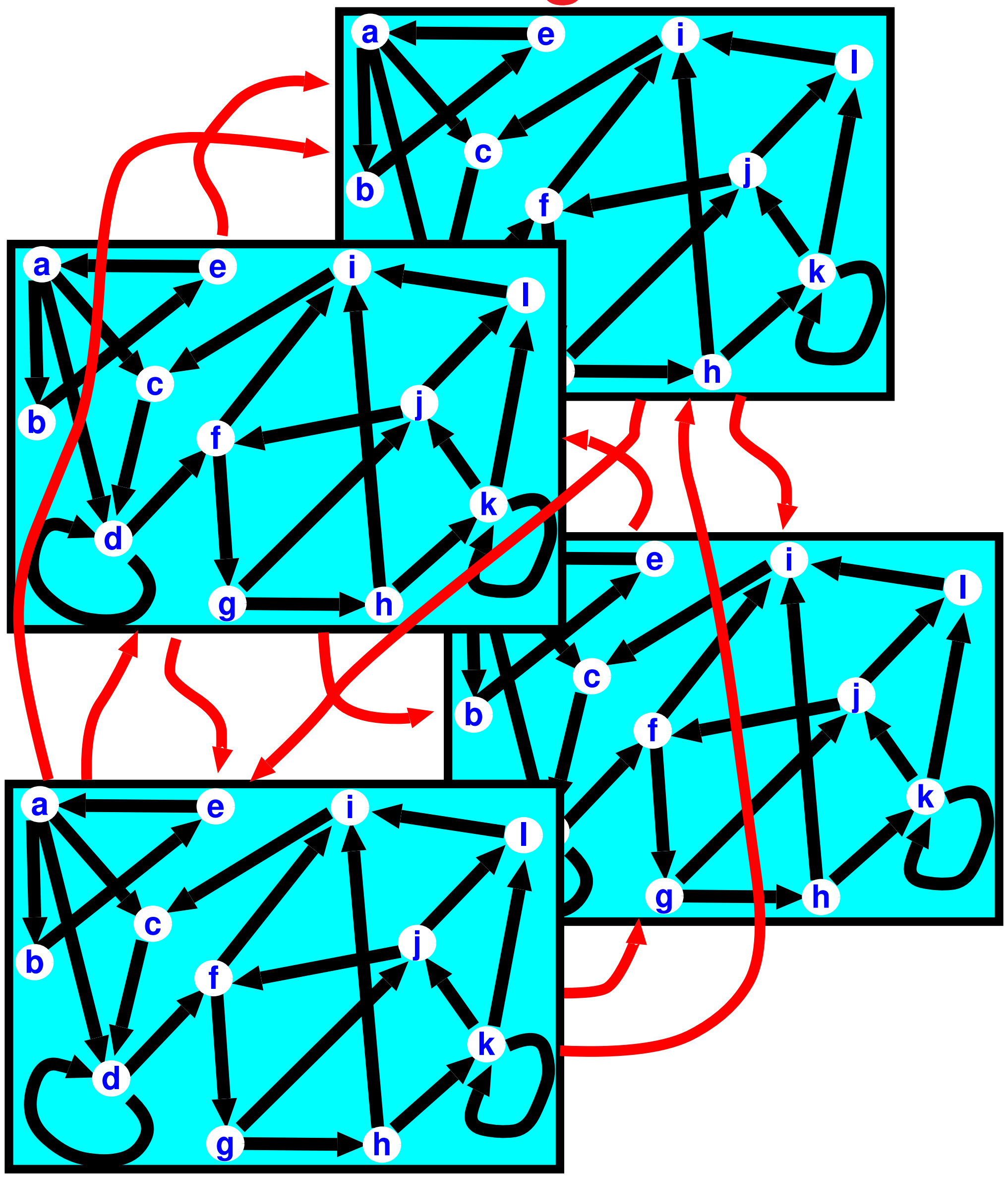
NOTE: functionally distinct sub-systems do not necessarily map onto physically separable sub-systems.

VMF and State Transition Diagrams

State transition diagrams and flow charts are possible only for systems that have "atomic states" (indivisible states)

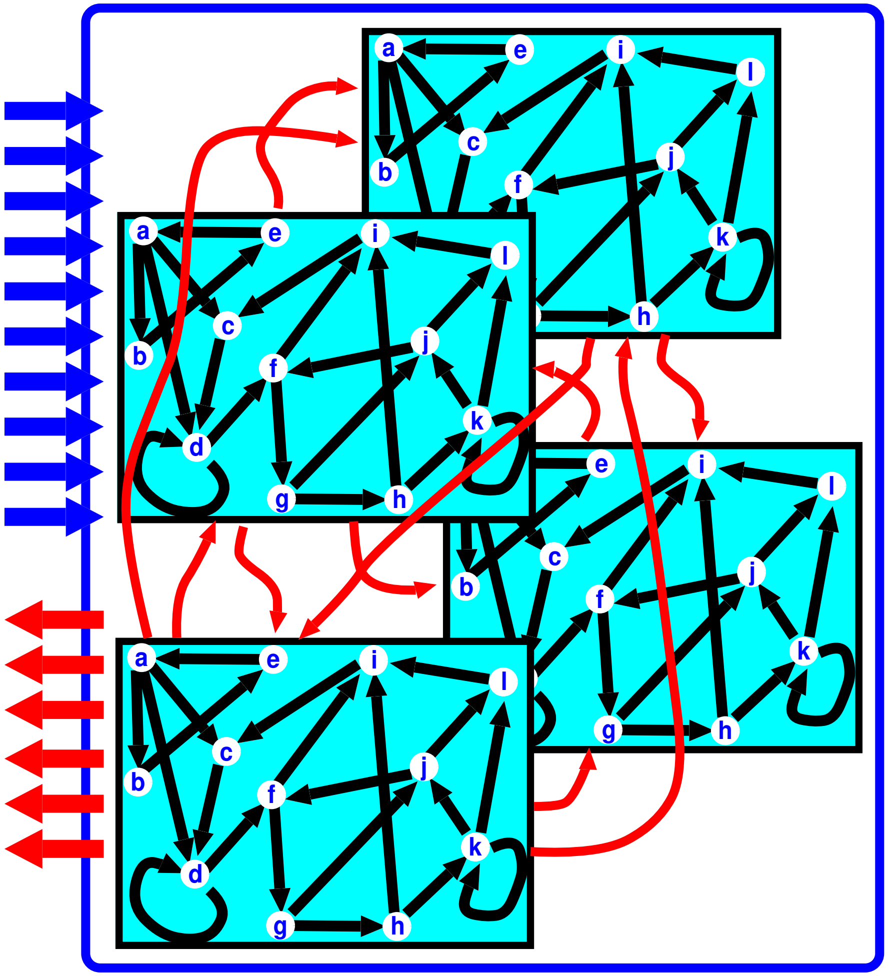
The sorts of virtual machines we are considering include multiple different co-existing virtual machines running in parallel (with their own state transitions) each with its own inputs and outputs, most of which are not connected to sensors and motors of the whole organism (or robot) but only to other such subsystems.

If there are multiple concurrently active subsystems that are not synchronised, a state transition diagram for the whole system is not possible.



Virtual machine Functionalism

We assume a kind of functionalism that allows many virtual machine components to co-exist and interact, including some that observe others, all within one agent.



Death of the Turing test

When the Input/Output bandwidth of the system is too low to reveal everything going on internally, there may be real, causally efficacious events and processes (including virtual machine events and processes) that cannot be directly observed.

> Even opening up the system may not make it easy to observe the VM events and processes (decompiling can be too hard).

If the links between systems can be turned on and off by internal processes, then during some states:

> some of the subsystems may not have any causal influence on outputs.

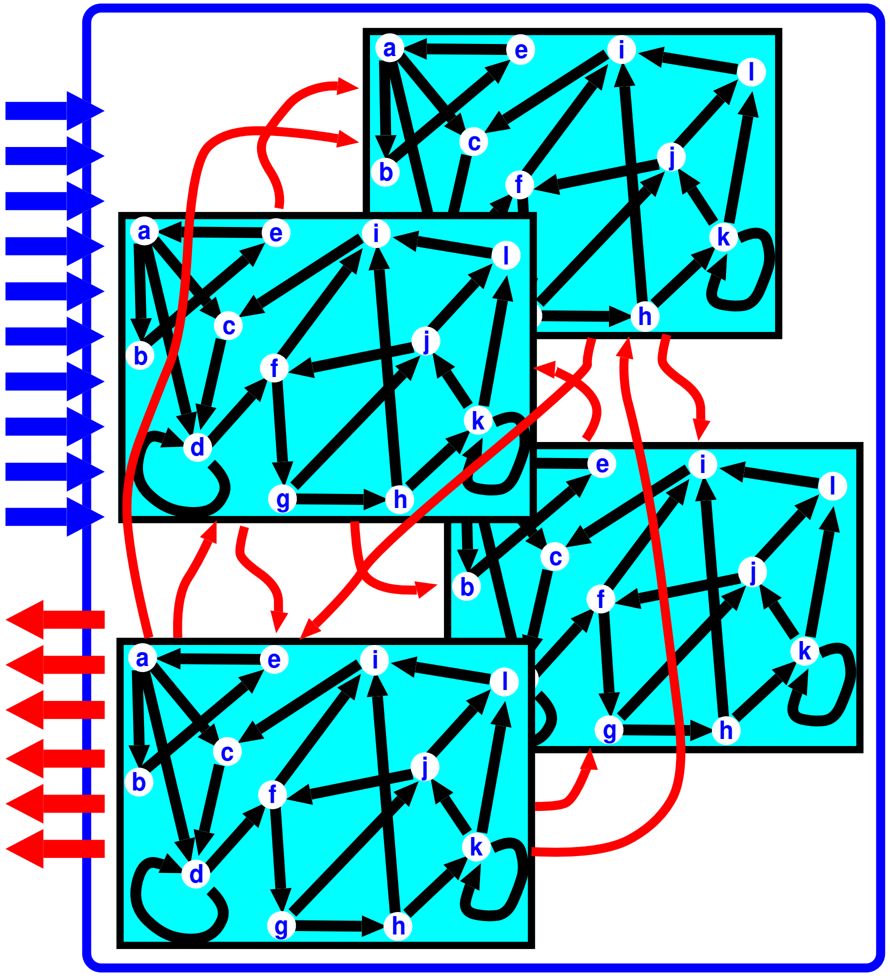
They still exist and can include internal causal interactions.

As theorists and designers we wish to be able to explore such systems and understand their implications, their strengths their weaknesses. (In the footsteps of evolution).

More on VM Functionalism

Instead of a single (atomic) state which switches when some input is received, a virtual machine can include many sub-systems with their own states and state transitions going on concurrently, some of them providing inputs to others.

- The different states may change on different time scales: some change very rapidly others very slowly, if at all.
- They can vary in their granularity: some sub-systems may be able to be only in one of a few states, whereas others can switch between vast numbers of possible states (like a computer's virtual memory).
- Some may change continuously, others only in discrete steps.



Some sub-processes may be directly connected to sensors and effectors, whereas others have no direct connections to inputs and outputs and may only be affected very indirectly by sensors or affect motors only very indirectly (if at all!).

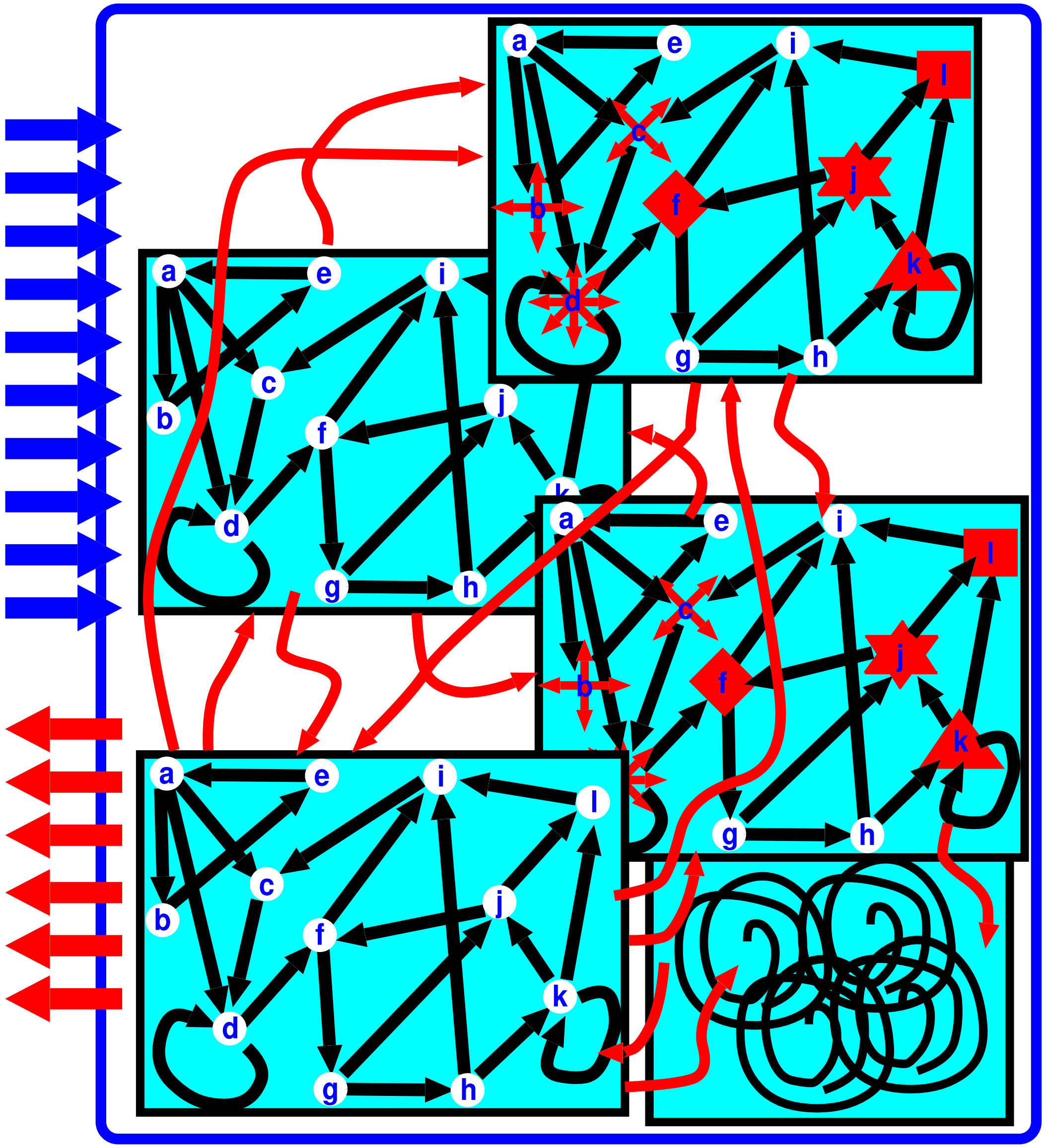
The previous picture is misleading

Because it suggests that the total state is made up of a fixed number of discretely varying sub-states:

We also need to allow systems that can grow structures whose complexity varies over time, as crudely indicated on the right, e.g. trees, networks, algorithms, plans, thoughts, etc.

And systems that can change continuously, such as many physicists and control engineers have studied for many years, as crudely indicated bottom right e.g. for controlling movements.

The label ‘dynamical system’ should be applicable to all these types of sub-system and to complex systems composed of them.



More features of VM's allowed by VMF

- Many forms of information may flow between subsystems, including
  - control information
  - factual information
  - questions,
    and
  - replies to requests, etc.
- Sub-systems can have functions that are hard (or impossible) to define in terms of input-output relations of the total system.

  A deep problem for psychology
- Some systems may be either totally disconnected or mostly disconnected from external connections.

  (could be an evolutionary accident)
- Sub-systems may operate on different time-scales: changing fast or slowly, discretely or continuously.

See A.Sloman (1993) The Mind as a Control System in *Philosophy and the Cognitive Sciences* Eds. C. Hookway & D. Peterson, pp. 69–110, CUP. Also at http://www.cs.bham.ac.uk/research/cogaff/

Information-structures of many kinds

It is also the case that individual sub-systems need not have atomic, indivisible states.

Information-bearing components of virtual machines may include

- Numeric values
- Vectors of values
- Trees and networks with changing structure and contents, e.g. topological maps,
- Logical formulae.
- Rule-systems.
- Maps, images and other “spatially” organised structures
- Activation levels in neural nets
- Other things as yet unknown

For instance, during the process of constructing a sentence various fragmentary representations of meaning may be constructed.

Likewise partially grown plans may exist.

There will also be representations of ontologies that grow during learning.

Causal laws for virtual machine states

An agent A can have changing numbers of co-existing sub-states, S1, S2, each distinguished by its causal connections and its laws of behaviour.

If A is in sub-state S, and simultaneously in various other sub-states, then

- ***if the sub-system of A concerned with S receives inputs I_1, I_2,... from other sub-systems or from the environment, and***
- ***if sub-states S_k, S_l, ... exist***
- ***then***
  - ***S will cause output O_1 to the sub-system concerned with state S_m***
  - ***S will cause output O_2 to the sub-system concerned with state S_n***
  - ***......***
  - ***and possibly other outputs to the environment (or to motors), and***
  - ***S will cause itself to be replaced by state S_2***
    where S_2 may differ in complexity from S_1
    (items may be destroyed or created: leading to a change in the number of coexisting sub-states)

NOTE: this formulation does not do justice to the rich variety of virtual machine processes that can be specified in computer programming languages. We are still discovering new types of processes that can be made to occur in virtual machines.

Varieties of causal interactions

Causal interactions within VMs may differ in ways not yet mentioned

- In some cases the causal interactions may be probabilistic rather than deterministic, e.g., if part of the context that determines effects of a sub-state consists of random noise or perturbations in lower levels of the system.
- In some cases the sub-states, their inputs and outputs vary continuously, whereas in others they vary discontinuously (discretely).
- Changes may be synchronised (or partially synchronised) or asynchronous.
- Individual sub-states may or may not be connected to external transducers.
- Some causal interactions simply involve quantitative effects, e.g. initiation, termination, excitation or inhibition, whereas others are far more complex and involve structural changes, e.g. transmission of structured information from one sub-system to another.
- Some sub-states may change in complexity as new parts or links are added to complex objects, e.g. creating a sentence, a sonnet, a proof or a plan in your mind.

VMF and Architectures

A virtual machine typically has an architecture: it is made up of interacting parts, which may themselves have architectures.

- **This kind of functionalist analysis of mental states is consistent with the recent tendency in AI to replace discussion of mere *algorithms* with discussion of *architectures* – in which several co-existing sub-systems can interact, perhaps running different algorithms at the same time, (e.g. minsky87, brooks86, sloman78).**
- **Likewise, many software engineers design, implement and maintain virtual machines that have many concurrently active sub-systems with independently varying sub-states.**
- **A running operating system like Solaris or Linux is a virtual machine that typically has many concurrently active components.**
- **New components can be created, and old ones may die, or be duplicated. Some enduring components may themselves have components that change.**
- **The internet is another, more complex, example.**

It does not appear that most philosophers who discuss functionalism, take explicit account of the possibility of virtual machine functionalism of the sort described here, even though most software engineers would find it obvious.

Virtual machines and mental processes

- Strong AI aims not merely to replicate the input-output behaviours of a certain kind of mind but actually to replicate the internal processes, That requires making sure that we know not merely what the processes are that normally occur in such a mind, but what the causal relationships are.
- That means knowing how various possible changes in certain internal structures and processes would affect other structures and processes even if normally those changes do not occur.
- I.e. replicating mental processes in virtual machines requires us to know a great deal about the causal laws and true counter-factual conditionals ("what would have happened if") that hold for the interactions in the system being replicated.
- Only then can we ask whether the artificially generated virtual machine truly replicates the original processes. But finding out what those laws are may be very difficult, and investigating some "what if" questions could be unethical!

Do Turing machines suffice?

It is not obvious that every collection of causal relations within human mental processes can be replicated by suitable processes running in a physical TM, since

- **It may be impossible to produce a TM implementation that supports the same set of counterfactual conditionals as some other implementation in which the higher level rules and interactions are more directly supported by the hardware.**
- **E.g. a neural net that is simulated on a serial machine typically goes through vast numbers of states which cannot occur in a parallel implementation where the nodes change state simultaneously, constraining causal interactions.**

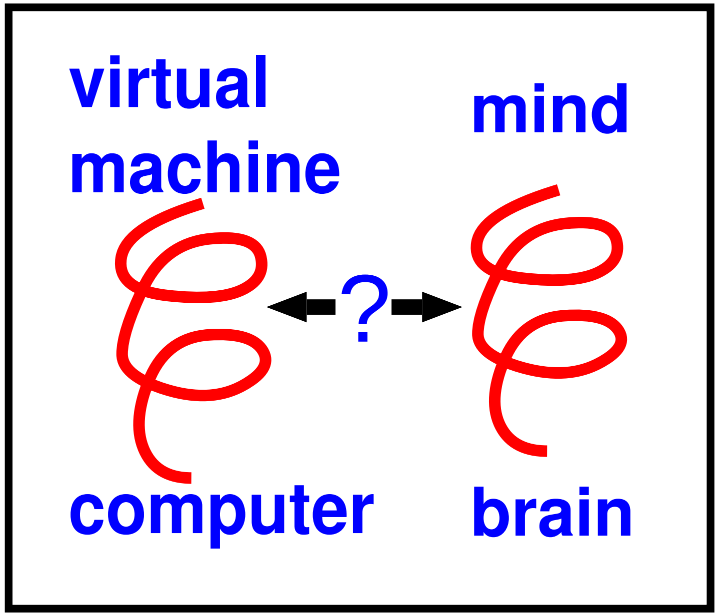
See

A. Sloman, 2002, The irrelevance of Turing machines to AI, in *Computationalism: New Directions*, Ed. M. Scheutz, MIT Press,

(Available at http://www.cs.bham.ac.uk/research/cogaff/),

‘Emergence’ need not be a bad word

People who have noticed the need for pluralist ontologies often talk about ‘emergent’ phenomena.
But the word has a bad reputation, associated with vitalist theories, sloppy thinking, wishful thinking, etc.



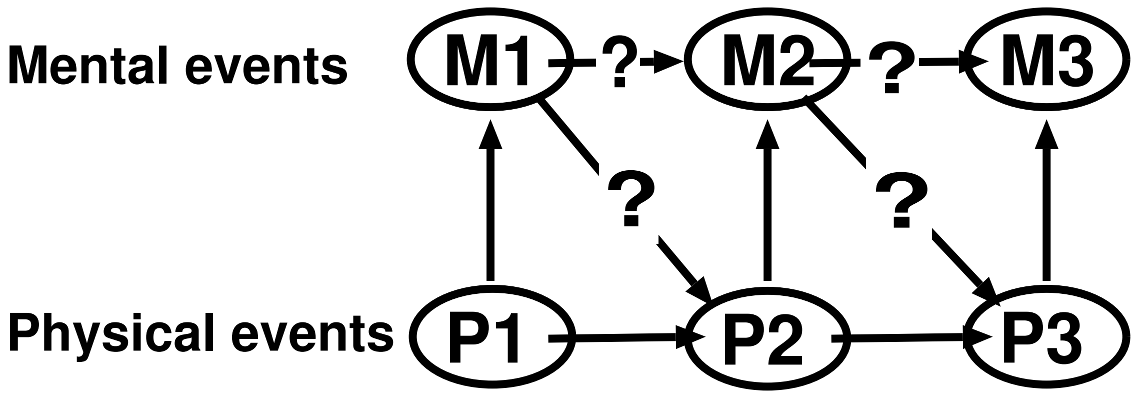
My claim: if we look closely at the kinds of ‘emergence’ found in virtual machines in computers, where we know a lot about how they work (because we designed them and can debug them, etc), then we’ll be better able to go on to try to understand the more complex and obscure cases, e.g. mind/brain relations.

Engineers discussing implementation of VMs in computers and philosophers discussing supervenience of minds on brains are talking about the same ‘emergence’ relationship — but they know different things about it.

Must non-physical events be epiphenomenal?

Many cannot believe that non-physical events can be causes.

- Consider a sequence of virtual machine events or states M1, M2, etc. implemented in a physical system with events or states P1, P2, ...



- If P2 is caused by its physical precursor, P1, that seems to imply that P2 cannot be caused by M1, and likewise M2 cannot cause P3.

  Moreover, if P2 suffices for M2 then M2 is also caused by P1, and cannot be caused by M1. Likewise neither P3 nor M3 can be caused by M2.

- So the VM events cannot cause either their physical or their non-physical successors.
- This would rule out all the causal relationships represented by arrows with question marks, leaving the M events as epiphenomenal.

The flaw in the reasoning?

THIS IS HOW THE ARGUMENT GOES:

- **Premiss 1: physical events are physically determined**

  E.g. everything that happens in an electronic circuit, if it can be explained at all by causes, can be fully explained according to the laws of physics: no non-physical mechanisms are needed (though some events may be inexplicable, according to quantum physics)

- **Premiss 2: physical determinism implies that physics is 'causally closed' backwards**

  I.e. if all caused events have physical causes, then nothing else can cause them: any other causes will be *redundant*.

- **Therefore: no non-physical events (e.g VM events) can cause physical events**

  E.g. our thoughts, desires, emotions, etc. cannot cause our actions. And similarly poverty cannot cause crime, national pride cannot cause wars, and computational events cannot cause a plane to crash, etc.

ONE OF THE PREMISSES IS INCORRECT. WHICH?

It's Premiss 2

Some people think the flaw is in the first premiss:

i.e. they assume that there are some physical events that have no *physical* causes but have some other kind of cause that operates independently of physics,
e.g. they think a spiritual or mental event that has no physical causes can cause physical events — 'acts of will' thought to fill gaps in physical causality.

The real flaw is in the second conjunct:

i.e. the assumption that determinism implies that physics is 'causally closed' backwards.

Examples given previously show that many of our common-sense ways of thinking and reasoning contradict that assumption.

Explaining exactly what is wrong with it requires unravelling the complex relationships between statements about causation and counterfactual conditional statements.

A sketch of a partial explanation can be found in the last part of this online tutorial, on philosophy of AI: http://www.cs.bham.ac.uk/˜axs/ijcai01

‘Emergent’ non-physical causes are possible

Problems with the ‘monistic’, ‘reductionist’, physicalist view that non-physical events are epiphenomenal:

- It presupposes a layered view of reality with a well-defined ontological bottom level. IS THERE ANY SUCH BOTTOM LEVEL?
- There are deep unsolved problems about which level is supposed to be the real physical level, or whether several are.
- It renders inaccurate or misleading much of our indispensable ordinary and scientific discourse, e.g.
  - Was it the government’s policies that caused the depression or would it have happened no matter which party was in power?
  - Your anger made me frightened.
  - Changes in a biological niche can cause changes in the spread of genes in a species.
  - Information about Diana’s death spread rapidly round the globe, causing many changes in TV schedules and news broadcasts, much sorrow, and many public demonstrations.

Identity theories

Identity theorists attempt to retain VM events as causes, while retaining physical events and states as the only causes.

The “identity theory” states that VM events can be causes because every VM event is just a physical event in the physical machine and since PM events can be causes, the VM events that are identical with them can also be causes.

However

- this identity theory does not explain anything deep, such as why not all physical configurations produce mental events;
- it contradicts the asymmetry in the realisation/supervenience relation;

  A running chess virtual machine is realised in and supervenient on the physical processes in the host computer but the physical processes are neither realised in nor supervenient on the chess processes: this lack of symmetry is incompatible with “identity” as normally understood.
- One manifestation of the difference is that VM events and PM events enter into different kinds of explanations, using different sorts of generalisations with different practical applications.

  E.g. understanding the VM event produced by a bug in a program, (for instance failing to distinguish two types of conditions in a ruleset) enables one to alter the program code (replace one rule with two, for different cases), and this repair will generalise across different running VMs using the same program on different kins of physical machines.

Against epiphenomenalism

- The argument that virtual machine events cannot have causal powers ignores how actual implementations of virtual machines work, and the ways in which they produce the causal powers, on which so much of our life and work increasingly depend.

  More and more control systems depend on virtual machines that process information and take decisions, e.g. controlling chemical plants, flying highly unstable aircraft.

- There is much more that is intuitively understood by engineers which has not yet been clearly articulated and analysed.
- The people who use this kind of understanding, but cannot articulate it could be called craftsmen rather than engineers.

  This is a special case of the general fact that craft precedes science.

  Philosophers and psychologists need to learn the craft, and the underlying science, in order to avoid confusions and false assumptions.

A more general notion of supervenience

Philosophers normally explain supervenience as a relation between *properties*: e.g. a person's mental properties are said to supervene on his physical properties.

> '[...] supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respects, or that an object cannot alter in some mental respect without altering in some physical respect.'
>
> D. Davidson (1970), 'Mental Events', repr. in: *Essays on Action and Events* (OUP, 1980).

In contrast we are concerned with a relation between *ontologies* or parts of ontologies, not just properties.

> The cases we discuss involve not just one object with some (complex) property, but large numbers of abstract objects enduring over time, changing their properties and relations, and interacting with one another: e.g. data-structures in a virtual machine, or thoughts, desires, intentions, emotions, or social and political processes, all interacting causally.

A single object with a property that supervenes on some other property is just a special case.

We can generalise Davidson's idea:

> **An ontology supervenes on another ontology if there cannot be a change in the first ontology without a change in the second.**

Notions of Supervenience

We can distinguish at least the following varieties

- **property supervenience**
  (e.g. having a certain temperature supervenes on having molecules with a certain kinetic energy.)
- **pattern supervenience**
  (e.g., the supervenience of a rotating square on the pixel matrix of a computer screen, or supervenience of various horizontal, vertical and diagonal rows of dots on a rectangular array of dots.)
- **mereological, or agglomeration, supervenience**
  (e.g., possession of some feature by a whole as the result of a summation of features of parts, e.g. the supervenience of a pile with a certain mass on a collection grains of sand each with its own mass)
- **mechanism supervenience**
  (supervenience of a collection of interacting objects, states, events and processes on some lower level reality, e.g., the supervenience of a running operating system on the computer hardware – this type is required for intelligent control systems)

We are talking about mechanism supervenience.
The other kinds are not so closely related to implementation.

Virtual machine functionalism assumes mechanism supervenience is possible.

The Physical Realization Theory of mind: PRT

The PRT states:

The mental is realised, or fully grounded, in the physical

or put differently,

if a collection M of mental objects/properties/states/events exists they have to be "fully grounded" in some physical system.

To say

An ontology of type O1 is fully grounded in an ontology of type O2

means

For an instance Io1 of type O1 to exist, there must be one or more instances of type O2, Io21, Io22, Io23... (possible implementations of Io1) such that:

– The existence of any of those instances of O2 is sufficient for Io1 and all of its properties and internal and external causal relations to exist

– For Io1 to exist, at least one of the possible implementations, of type O2, must exist (i.e. no disembodied mental objects events, processes, etc.)

The instance of type O2 that realises Io1 need not be unique: multiple realisation is possible

All this fits the engineer's notion of implementation, though engineers know a lot more about the details of how various kinds of implementations work.

(They intuitively understand and make use of mechanism supervenience.)

NOTE

The physical realisation thesis is probably what Newell and Simon meant by their "physical symbol system" hypotheses.

Their terminology is very misleading because most of the symbols AI systems deal with are not *physical* symbols, but *symbols in virtual machines*.

They should have called it

the physically implemented virtual symbol system

hypothesis.

Realisation (grounding) entails supervenience

IF

M is fully grounded in P (as defined above),

THEN IT FOLLOWS LOGICALLY THAT

no feature, event, property or relation of M can change unless something changes in P,
since otherwise the original state of P was sufficient for the new feature,
yet the new feature did not appear.

Note: this would not be true if some changes in M are produced by spirits or souls that are not implemented in any physical systems.

PROOF:

If M is fully implemented in P, then every facet of M is explained by P.

If there's a change in M, that would introduce a new facet not explained by P (since P was not sufficient for it before the change).

Therefore: something must have changed in P which explains the change in M.

I.e.

Physical realisation entails supervenience.

Difference in physical machines does not imply difference in VMs, but difference in VMs implies physical differences.

(Actually we need to discuss more cases — another time.)

Surprising aspects of implementations.

- The relation may be "partial" (if considered at a particular point in time) if there are entities in the VM which do not correspond to physical parts.

  A partial implementation of a large array might have cells that will be created only if something needs to access their contents. A collection of theorems might exist *implicitly* in a mathematical virtual machine, but not be explicitly recorded until needed. It may be partial in another way if there are physical limitations that prevent some VM processes occurring, e.g. memory limits.
- A VM may contain a huge 'sparse array' with more items in it than there are electrons in the computer implementing it.

  (e.g., cells containing items with some computable or default value are not explicitly represented)
- Individual VM entities may map on to physical entities entities in different ways.

  (e.g., some VM list structures might be given distinct physical implementations, while others share physical memory locations because they have common 'tails'.)
- In a list-processing language like Lisp or Pop-11, there can be two lists each of which is an element of the other, whereas their physical counterparts cannot be so related.
- If the virtual machine uses an interpreted programming language, then the mapping between high level objects or events and physical entities is constantly being (re-)created by the interpreter.

  If the language is compiled then the causal powers of program code are different: the execution is more ballistic.
- A learning, self-optimising, machine may change the way its virtual entities are implemented, over time.

Disconnected virtual machine components

There are two interesting variants of VMF, restricted and unrestricted.

Restricted virtual machine functionalism **requires that every sub-state be causally connected, possibly indirectly, to inputs and outputs of the whole system A.**

Unrestricted **VMF does not require this.**

E.g. it allows a virtual machine to include a part that does nothing but play games of chess with itself.

More interestingly, a sub-mechanism may be causally disconnected some of the time and engaged at other times.

Causal connectivity to inputs/outputs is also not required by atomic state functionalism as normally conceived, since a finite state machine can, in principle, get into a catatonic state in which it merely cycles round various states forever, without producing any visible behaviour, no matter what inputs it gets.

A philosophical view very close to restricted VMF was put forward by Ryle (1949), (e.g., in the chapter on imagination), though he was widely misinterpreted as supporting a form of behaviourism.

The possibility of “causally disconnected” VMs explains a number of philosophical puzzles about consciousness.

(Explained in a forthcoming paper by Sloman and Chrisley in JCS.)

This undermines a number of common assumptions of psychological research: it is possible for mental states to exist that are not empirically detectable “from outside”.

Biological virtual machines

Biological evolution produced many kinds of information-processing machine.

Such machines are very different from matter-manipulating or energy-manipulating machines.

Many forms of consciousness (fish, fly, frog and adult human consciousness) are products of biological evolution (aided and abetted by individual development and social development).

Each form of consciousness required the evolution of appropriate information-processing capabilities of organisms (including possibly some virtual machines with partly or wholly “disconnected” components).

From that viewpoint, treating consciousness as one thing can be seen as analogous as treating motion as one thing, ignoring

- the huge variety of differences in what motion achieves in microbes, plants, insects, fishes, birds, mammals, etc.
- the huge variety of ways in which it is produced
- the huge variety of ways in which it relates to other things going on in the organism.

(These variations should not be thought of as continuous, or mere matters of degree: biological differences are inherently discrete.)

Understanding animal consciousness

If we wish to understand the biological phenomena as opposed to some theoretical philosophers' (or physicists') abstraction, we need to understand

- the varieties of information-processing mechanisms available,
- the ways they can be combined in complete functional architectures,
- the reasons why virtual machine architectures are relevant,
- the varieties of forms of consciousness that can arise in all these different architectures.

One important characteristic of the hypothesised human information processing architecture is a consequence of "co-evolution" of perceptual, action, and central sub-systems, with advances in each generating new requirements in the others and enabling new advances in the others.

Much of human (and animal) perceptual consciousness should be seen more as consciousness of affordances than as consciousness of things, properties and relationships.

Another consequence is that insofar as humans include different sub-systems each capable of supporting different kinds of consciousness (as they do in different organisms) it follows that humans have different kinds of consciousness.

Further details available here http://www.cs.bham.ac.uk/research/cogaff/talks/

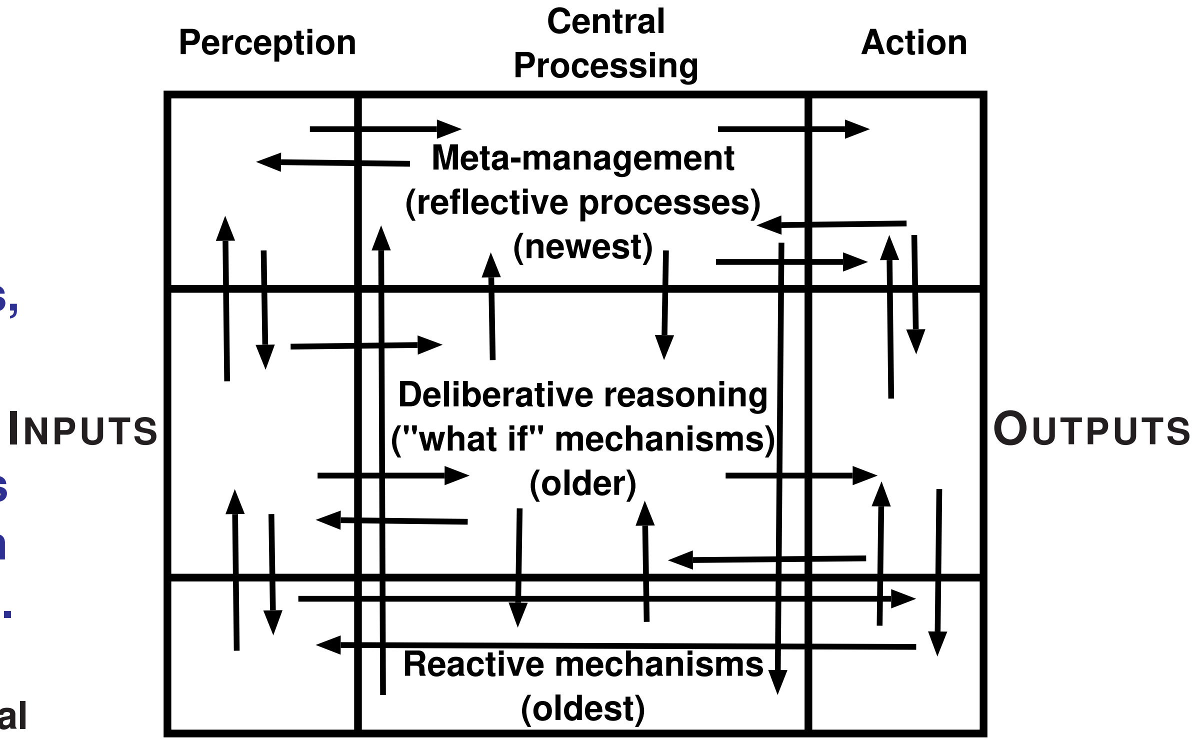
We need to understand the space of VM architectures

A sort of generative grammar for a class of architectures for integrated agents, perceiving and acting on: a complex and changing environment.

Different architectures include mechanisms in different subsets of the boxes, and different possible information links, different possible control relationships.

There are also differences in forms of representation and types of mechanisms.

There are many particular architectures that fit this general framework. E.g. it seems that insects have architectures containing only mechanisms in the reactive layer.



Not all possible information flows are shown!

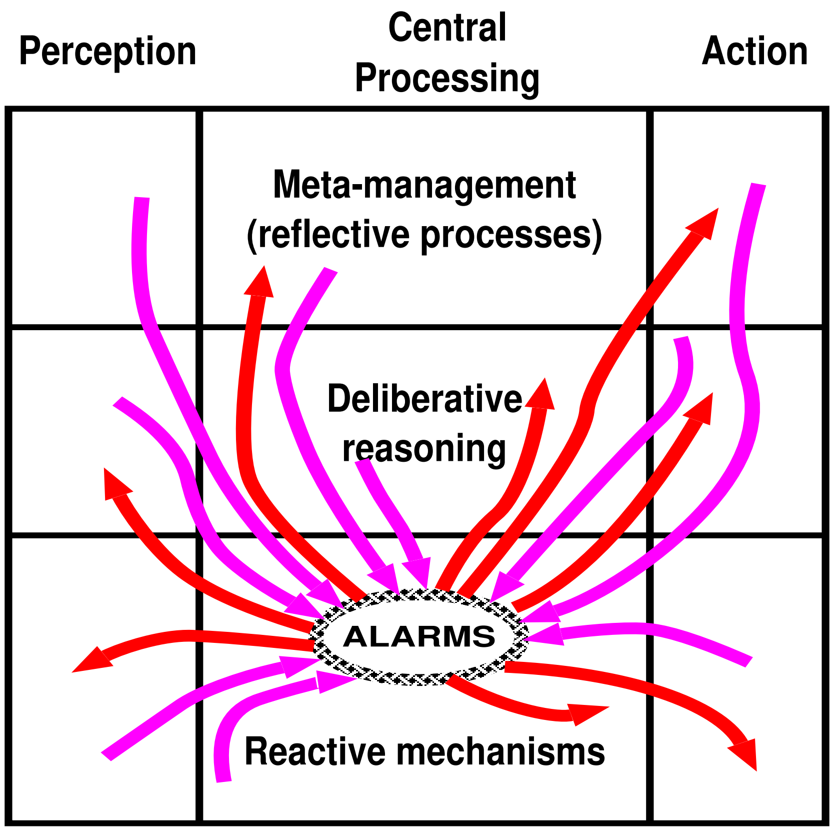
The Cognition and Affect papers and presentations explain this in more detail.

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

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

With alarm mechanisms

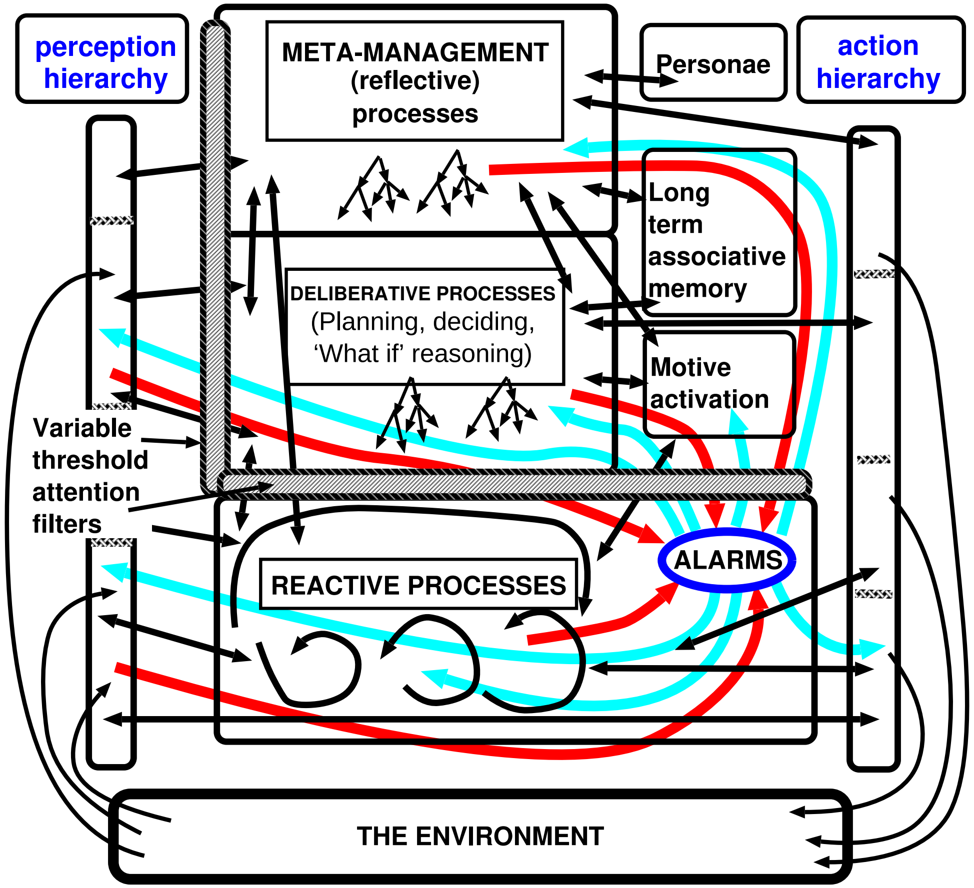
- Alarms allow rapid redirection of the whole system or specific parts of the system required for a particular task (e.g. blinking to protect eyes.)
- The alarms can include specialised learnt responses: switching modes of thinking after noticing a potential problem.
- E.g. doing mathematics, you suddenly notice a new opportunity and switch direction. Maybe this uses an evolved version of a very old alarm mechanism.
- The need for (POSSIBLY RAPID) pattern-directed re-direction by meta-management is often confused with the need for emotions e.g. by Damasio, et. al.



The architectural basis for a subset of emotions

Sketch of H-Cogaff, a possible architecture for human-like systems

All components shown operate concurrently.



An architecture involving huge numbers of counter-factual conditionals.

Explained in more detail here: http://www.cs.bham.ac.uk/research/cogaff/talks/#talk24

(And other talks in the same directory)

Conjectures: Part 1

- Something like the CogAff schema can enable our field to make better progress by providing a common ontology and vocabulary for comparing and contrasting architectures.
- Something like the H-Cogaff architecture, the conjectured human-like instance of CogAff, can accommodate many features of human mental functioning, including emotions and learning.

  It allows for

  – more varieties of emotions than are normally considered

  – diverse affective states, including desires, pleasures, pains, attitudes, moods, etc.

  – many varieties of learning and development, involving different parts of the architecture and new links between parts,

  – maybe even "qualia" solving old philosophical problems about consciousness.

But I don't have time to explain all this today.

Conjectures: Part 2

- Architecture-based concepts defined in terms of H-Cogaff, can refine and extend many of our *very confused* concepts,
  e.g. emotion, learning, consciousness (cluster concepts).

- Related Architecture-based concepts, using other instances of the CogAff framework, will be applicable to other organisms and machines,
  e.g. "emotions" in insects.

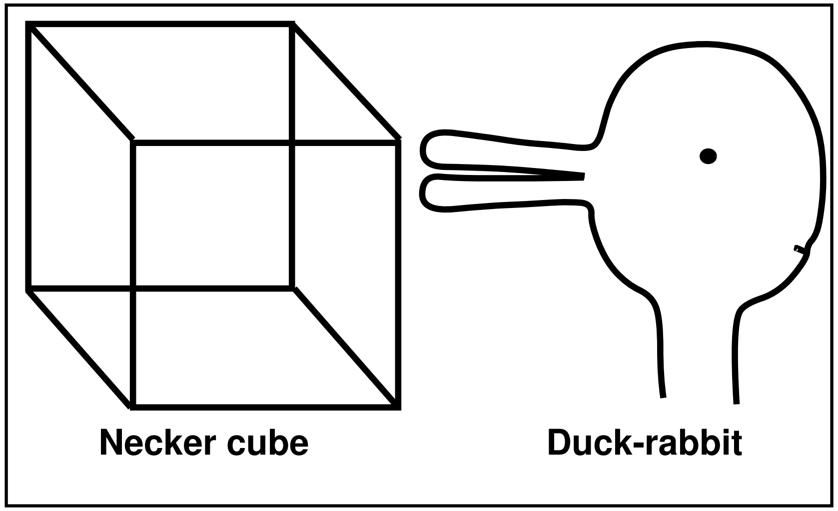
- Models based on H-Cogaff can play a useful role both for
  - applications of AI (e.g. digital entertainments, or in learning environments), and for
  - scientific theories about human minds and brains.

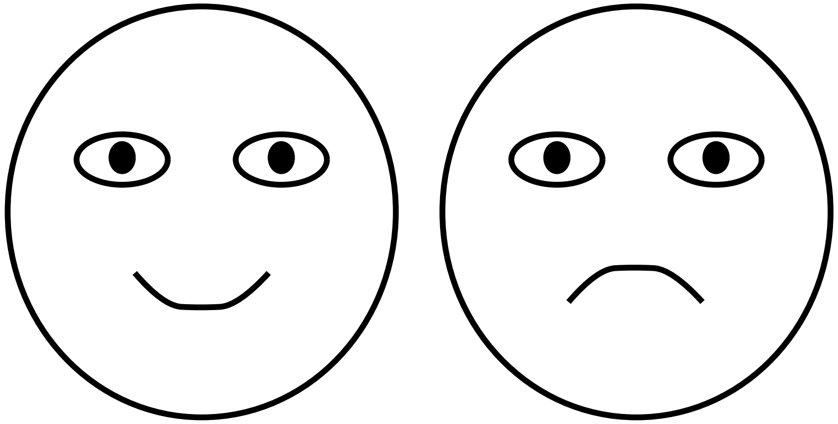
NOTE:

At present the specifications are extremely sketchy.
Some details could be filled in using known AI techniques and formalisms.
Most of it requires major future advances, some discussed in other slide presentations.
Vision, especially, is largely not understood: e.g. we don't know what its functions are.

I give some partial answers on the Cognition and Affect web site.
See http://www.cs.bham.ac.uk/research/cogaff/talks/

Ontologies in perceptual mechanisms





Seeing the switching Necker cube requires geometrical percepts.

Seeing the flipping duck-rabbit uses far more subtle and abstract percepts, going beyond geometric and physical properties.
(Contrast Marr's views on vision, and much AI vision research.)

Things we can see besides geometrical properties:

- **Which parts are ears, eyes, mouth, bill, etc.**
- **Which way something is facing**
- **Whether someone is happy, sad, angry, etc.**
- **Whether a painting is in the style of Picasso...**

This helps to explain why the functions of vision go beyond the detection of physical and geometric features of the environment.

A new kind of explanation?

Notice that we are not saying: This is what qualia are .... Instead, we offer (conjectured) *sufficient* conditions for an information processing system to go through the very same processes as led humans to start thinking about sensory qualia and also other kinds. It's a side-effect of sophisticated biological mechanisms.

A meta-management system could give an agent the ability to attend not only to what is perceived in the environment, but to also features of the *mode of perception* that are closely related to properties of intermediate sensory data-structures.

Thus you can attend not only to (a) the table and its fixed 3-D shape, but also to (b) the 2-D *appearance* of the table in which angles and relative lengths of lines change as you change your viewpoint (or the table is rotated). The appearance can also change as you squint, tap your eyeball, put on coloured spectacles, ...

This is exactly the sort of thing that led philosophers (and others) to think about qualia as something internal, non-physical, knowable only from inside, etc.

It suffices to explain the mental mechanisms that generate an interest in thinking and talking about qualia. Robots with our information processing architecture would do the same.

Towards a more specific account of qualia

Phenomena described by philosophers as “qualia” may be explained in terms of high level control mechanisms with the ability to switch attention from things in the environment to *internal* states and processes, including intermediate sensory datastructures in layered perceptual systems. These introspective mechanisms may explain a child’s ability to describe the location and quality of its pain to its mother, or an artist’s ability to depict how things look (as opposed to how they are). Software agents able to inform us (or other artificial agents) about their own internal states and processes may need similar architectural underpinnings for qualia.

From this standpoint, the evolution of qualia would not be a single event, but would involve a number of steps as more kinds of internal states and processes became accessible to more and more kinds of self-monitoring processes with different functions, e.g. requesting help from others or discovering useful generalisations about oneself. Such step-wise development may also occur within an individual.

A basis for sensory “qualia” can be found in self-monitoring mechanisms which give access to intermediate sensory information structures not normally attended to. Different kinds of sensory qualia would depend on different abstraction layers within perceptual mechanisms. This provides self-knowledge in a manner which is distinct from normal perception providing knowledge about the environment. Such “meta-management” capabilities can provide other sorts of qualia related to thinking processes, deliberation, desires, etc.

This internal monitoring of intermediate sensory stores happens for example, when you learn to see that besides the rectangular shape of the table in front of you there is also a skewed shape (a parallelogram) which is not out there but is part of an intermediate representation within you, determined by the structure of the viewpoint relative optic array: i.e. it changes shape as you move. Without this internal monitoring capability, “realistic” 2-D painting or drawing of 3-D objects would be impossible.

The biological functions of the “qualia” are real, but too complex to describe in detail here. E.g.

they can play a role both in individual learning, and also in communications where one agent tries to help another “debug” an internal process:

“Look just a teeny bit to the left of the large green triangle, dear.... then you’ll see it”

where there is no green triangle out there, but from the current viewpoint there’s a triangular region caused by e.g. the way two sloping walls and a floor bound a visible part of a rectangular lawn).

On this model there are many different kinds of qualia (contents of self-monitoring states), and they are to be explained at a functional level in terms of the architecture that makes them possible, and at a physical or physiological level in terms of the mechanisms used to implement the architecture, which may be different in different organisms or machines.

Robots with similar meta-management capabilities are likely to invent philosophical problems about qualia – and may wonder whether humans have them.

Roughly: qualia are what humans or future human-like robots refer to when referring to the objects of internal self observation.

Different sorts of qualia are connected with different contents of internal perception.

A special type of case arises out of the use of self-organising classifiers to categorise internal states (e.g. Kohonen nets).

A predicate produced in that way implicitly always refers to the system in which it is being used (causal indexicality: John Campbell). So qualia descriptors used by different individuals cannot be compared: in that sense qualia descriptors are inherently *private*.

(Like points of space referred to in different inertial frames.)

These topics are discussed more fully in A. Sloman and R.L. Chrisley, (2003), Virtual machines and consciousness, in *Journal of Consciousness Studies*, 10, 4-5
and older papers at the Cognition and Affect web site http://www.cs.bham.ac.uk/research/cogaff/

Meta-management mechanisms

- They can monitor, categorise, evaluate, and (to some extent) control other internal processes – e.g. some deliberative processes, or some perceptual processes. (See Barkleys's 1997 book on ADHD)
- This includes control of attention, control of thought processes. (Control which is lost in tertiary emotions.)
- Both monitoring and control depend on special purpose low level support for new architectural information pathways.
- They can vary widely in sophistication, e.g. depending on social learning.
- They require concepts and formalisms suited to self-description, self-evaluation
- They support a form of internal perception which, like all perception, may be incomplete or inaccurate, though generally adequate for their functional role.
- The concepts and formalisms may be usable in characterising the mental states of others also.
- Different meta-management control regimes may be learnt for different contexts (different socially determined "personae").
- Evolution of sensory qualia: occurs when it is useful for meta-management to look inside intermediate levels of perceptual processing (why?).
- If meta-management mechanisms are damaged, blind-sight phenomena may occur. (Experiments requiring subjects to *report* what they see typically use the meta-management layer! What's happening in other layers may be unnoticed.)

Why perception mechanisms are layered

Perception uses different layers to provide different sorts of information for different parts of the central system.

Parts of the perceptual hierarchy and parts of the central system co-evolve: each is part of the niche for the other.

The “highest” perceptual layers involve the use of the ontology for self-description developed in the meta-management layer. The same ontology can be used to describe other information-processing systems.

Hence our ability to see another as happy, sad, bored, attentive, etc.

Similar comments can be made about layers in the action system.

(For more on this see talks on vision here:
http://www.cs.bham.ac.uk/research/cogaff/talks/)

A type of TM-based attack on AI: the "absurd-implementation" attack

Various philosophers (e.g. Ned Block, Selmer Bringsjord, John Searle) have attacked a 'strong' version of AI as follows.

1. Strong AI implies that there is some AI program AIP for a TM which, when run, necessarily creates consciousness, intelligence, emotions, etc.
2. How a TM is physically implemented does not matter.
3. So AIP could be run by having millions of Chinamen blindly cooperating in following some rules (like the transistors in a computer), or it could be run by John Searle blindly following the rules.
4. But it is "obvious" (why??) that such implementations of a TM running AIP would not have consciousness, intelligence, emotions, etc.
5. Therefore strong AI is false.

Replies to “absurd-implementation” arguments

- Premiss 1 is false.
- The assumption that how something is implemented does not matter needs to be qualified.

  For engineering purposes it does matter.
  The implementations proposed in these anti-AI arguments would not satisfy the same collection of counterfactual conditional statements as a more conventional implementation (e.g. on computers).

  E.g. it is not true of a computer that one of its components can decide to stop cooperating, or maliciously give a wrong answer.
- In any case it is not clear that Searle or any other human can mentally simulate an architecture with the huge amount of concurrency in H-Cogaff, with several individual components interacting causally with the environment.

  (This is especially true of the full richness of visual perception.)

  Thus these implementations would not have the causal powers required for a mind with information autonomy (defined below). They are red herrings.

See

http://www.cs.bham.ac.uk/research/cogaff/0-INDEX00-02.html#70

http://www.cs.bham.ac.uk/research/cogaff/0-INDEX81-95.html#12

Machines can vary in what they operate on, use, or manipulate:

1. Matter
2. Energy
3. Information

- Scientists and engineers have built and studied the first two types of machines for centuries. Newton provided the first deep systematisation of this knowledge.
- Until recently we have designed and built only very primitive machines of the third type, and our understanding of those machines is still limited.
- Evolution 'designed' and built a fantastic variety of machines of all three types – with amazing versatility and power long before human scientists and engineers ever began to think about them.
  - Biological organisms are information-processing machines, but vary enormously in their information-processing capabilities.
  - There are myriad biological niches supporting enormously varied designs, with many trade-offs that we do not yet understand (e.g. trade-offs between cheapness and sophistication of individuals).
  - The vast majority of organisms have special-purpose information-processing mechanisms with nothing remotely like the abstractness and generality of TMs
  - A tiny subset of species (including humans) developed more abstract, more general, more powerful systems. Turing's ideas about TMs were derived from his intuitions about this aspect of human minds. But at best that's a small part of a human mind.

Understanding all this requires us to think more about architectures than about algorithms.
See: http://www.cs.bham.ac.uk/research/cogaff/talks/

Another kind of variation: autonomy

Machines can be more or less autonomous in various ways.

We can distinguish two main kinds of autonomy: energy autonomy and information autonomy, and sub-cases of the latter.

- **A machine may lack energy autonomy, e.g. because a person or other external agency has to turn a handle to provide energy, while having information autonomy because the machine determines how the energy is used.**

  Example: a music box, automatic loom, or mechanical calculator which requires a person to turn a handle to provide energy, while the machine decides what to do at each step.

- **A machine may lack information autonomy because a human or other external agency has to continually specify what it should do, even if it has energy autonomy.**

  Example: Someone driving a car is constantly providing control information through the use of steering wheel, brake pedal, accelerator, etc. The human does not provide energy, unlike a human riding a bicycle who provides both control information and energy.

- **Biological organisms typically exhibit both energy autonomy (except when energy levels are low and food is consumed) and information autonomy, of different kinds.**

Variations in information autonomy

Information autonomy includes different sorts of information that are useful to the organism or machine

- **Factual information**, which can be stored internally and used immediately, or later, including:
  - Information about particulars (objects, events, places, routes, times)
  - General information, universal or probabilistic, e.g. unsupported objects fall.
    Nothing is assumed about the form in which such information is stored: we still don't know much about varieties of forms of representation and their trade-offs.
- **Control information**, about what to do, including:
  - goals: immediate, future, or generic
  - plans: specific or generic ways of achieving goals
  - skills ('compiled plans').

  These may be built in (innate), learnt over a long period of training (e.g. in altricial species) or generated when needed in response to circumstances – and maybe then learnt.

Organisms vary in their information autonomy. However even when the vast majority of actions are products of internal information processing, the decisions normally make use of both currently sensed and previously acquired factual or control information developed by sensing or interacting with the environment.

Thus the information-processing system is *organism plus environment*, and unless the environment can be proved to be equivalent to a TM, theorems about TMs prove nothing about limits of biological organisms or biologically inspired AI systems.

Computers and brains

It used to be claimed that brains and computers had a lot in common.

Then it became fashionable to point to the differences: another way of attacking AI.

We need to understand that there are different levels of description and there may be deep similarities at some levels of description and huge differences at other levels.

But computers can also have huge differences at certain levels: e.g. compare current computers and the 1950s models.

Some high level similarities between brains and computers:

- **Connections to sensors and motors**
- **Storing, using transforming information.**
- **Multiple forms of processing in parallel – many asynchronously.**
- **Huge state-space provided by using large numbers of independently switchable units, and an even larger space of possible trajectories.**
- **Virtual and physical machines running at the same time with mutual support.**
- **Both long term information stores and currently active rapidly changing information stores (e.g. sensor driven temporary stores).**
- **Some of what the processes operate on are external and some are internal (e.g. learning, planning, correcting mistaken beliefs, choosing a goal).**

AI and mathematics

I suspect that an even more important development lies in the future

As we continue to develop AI theories about the architecture of mind, especially meta-management mechanisms, and apply them to modelling development of mathematical understanding in children, we shall find that many of our ideas about mathematics and how to teach mathematics are altered in profound ways.

Conclusion

No presentation like this can settle the questions.

But I hope I have at least done something useful

- by making clearer some of the features of virtual machines that are relevant to AI (the live presentation includes a demonstration of a working virtual machine with causal powers!),
- explaining how working, causally efficacious, virtual machines, despite being abstract objects, are quite different from the abstractions of pure mathematics, which have no causal interactions

There are important unsolved problems about the requirements for such virtual machines, including the forms of representation they may need, the varieties of ways they can be combined in complex architectures, and whether formally equivalent implementations are causally equivalent.

The best way forward is not to spend much time discussing these rather abstract questions, but to focus most effort on designing and testing much more detailed ideas about actual working systems with human-like capabilities.

As a side-effect, that will give us new ways of thinking about these debates.

THANKS

- To Matthias Scheutz, Ron Chrisley and Jeff Dalton for useful discussions on the specifics of this paper.
- To Margaret Boden, Stan Franklin, Marvin Minsky and Pat Hayes for much interesting discussion over many years related to this topic.
- To colleagues and students in the Birmingham Cognition and Affect project: http://www.cs.bham.ac.uk/˜axs/cogaff.html

  With papers and slide presentations here:

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

  http://www.cs.bham.ac.uk/research/cogaff/talks/
- To the Leverhulme Trust for research support
- To Dale Lichtblau for reporting errors in an earlier version.

To the developers of Linux and other free, portable, reliable, software systems, e.g. Latex, Tgif, xdvi, ghostscript, Poplog/Pop-11, etc.

No Microsoft software required