Talk presented at University of Bielefeld 10th Oct 2007
As part of the opening ceremony of COR-Lab
http://www.cor-lab.de/eng/

Why robot designers need to be philosophers

and vice versa

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

These slides are in my 'talks' directory:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#bielefeld

Philosophy overlaps with Artificial Intelligence (including robotics) much more than most people realise.

1. Doing philosophy can help designers of intelligent systems be clear about the goals they are aiming for, and the criteria by which their work should be evaluated.

2. Learning about designs for intelligent information processing systems helps to shed light on some old philosophical problems, e.g.

- problems about the relationships between mind and body
- problems about free will and determinism

And several more.

Some tutorial presentations

The following provide more information about the overlap between philosophy and AI:

- Talk 10: What is Artificial Intelligence?
  `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#whatsai`
- Talk 13: Artificial Intelligence and Philosophy
  `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#aiandphil`
- Others: `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/`

Also: A. Sloman, The Computer Revolution in Philosophy: Philosophy science and models of mind (1978)
Online here `http://www.cs.bham.ac.uk/research/projects/cogaff/crp`

Obvious Motivations for studying AI/Robotics

Motivations can be

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

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

- Solving existing practical problems

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

- Solving anticipated practical problems

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

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

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

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

Less obvious Motivations for studying AI/Robotics

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

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

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

  E.g. questions about information processing architectures, forms of representation, mechanisms.
  ESPECIALLY Questions about varieties of virtual machines, what they are useful for, and how they can be implemented – in brains or other kinds of physical machines.

- **Making progress with old philosophical problems**

  by providing new conceptual tools

  for articulating the questions and previously unthought of answers
  Including tools for demonstrating and testing philosophical theories

  Example:

  I originally got into AI because I wanted to show why Kant's philosophy of mathematics was correct and Hume's wrong – and eventually I realised that that goal required me to learn how to design a working mathematician – starting from a baby mathematician seeing shapes and learning to count!
  See chapter 8 of The Computer Revolution in Philosophy (1978)
  `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`
  Alas it proved much more difficult than I had anticipated – we still are not close! But see these two (a presentation and a paper, both written 2008): `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#math-robot` Could a Child Robot Grow Up To be A Mathematician And Philosopher?
  `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0804` Kantian Philosophy of Mathematics and Young Robots

How philosophy can contribute: consciousness

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

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

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

Example:

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

How philosophy can contribute: consciousness

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

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

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

Example:

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

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

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

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

Philosophical analysis can show

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

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

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

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

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

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

AI contributes to conceptual analysis in philosophy

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

  AI architectures can generate new "periodic tables" of types of mental processes.

  As explained in this introduction to logical geography *vs* logical topography:
  `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html`
- AI has already begun to revise our ontology for mental states and processes by showing us new, previously unimagined, subdivisions:

  e.g. different sorts of learning, different levels of control; different functions for perception; different sorts of processes related to our notion of "emotion".

Some virtual machine demos

The talk presented a live demo of the sort shown in video recordings here:

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

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

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

I feel glum because ...

I feel surprised because ...

I feel happy because ...

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

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

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

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

Virtual machines and an old philosophical problem:

What is the relation between mind and body?

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

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

We now show how to get some things clearer.

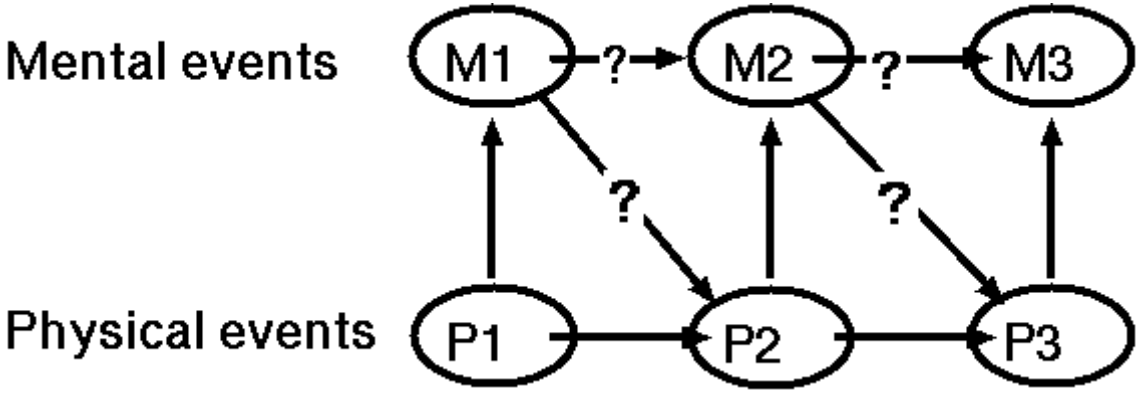
Supervenience and the mind-body relation

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

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

But there are many problems about that relationship: can mental process cause physical processes?

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

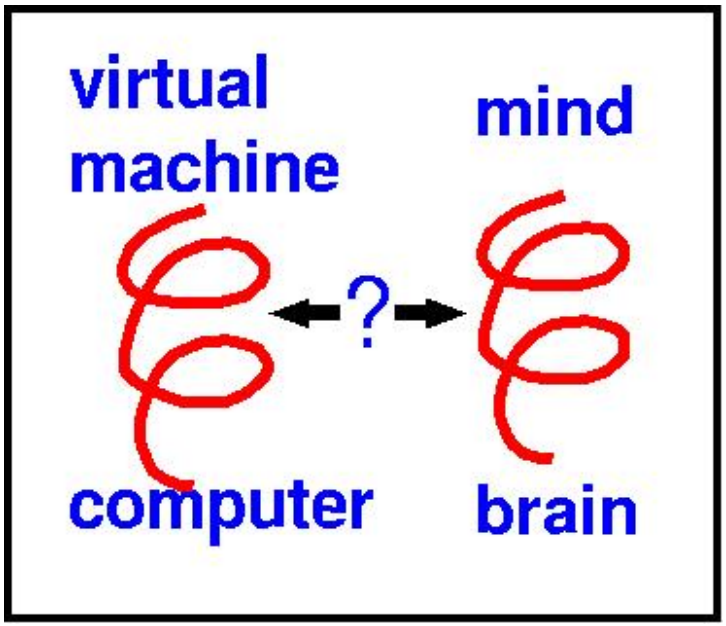


(Think of time going from left to right)

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

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

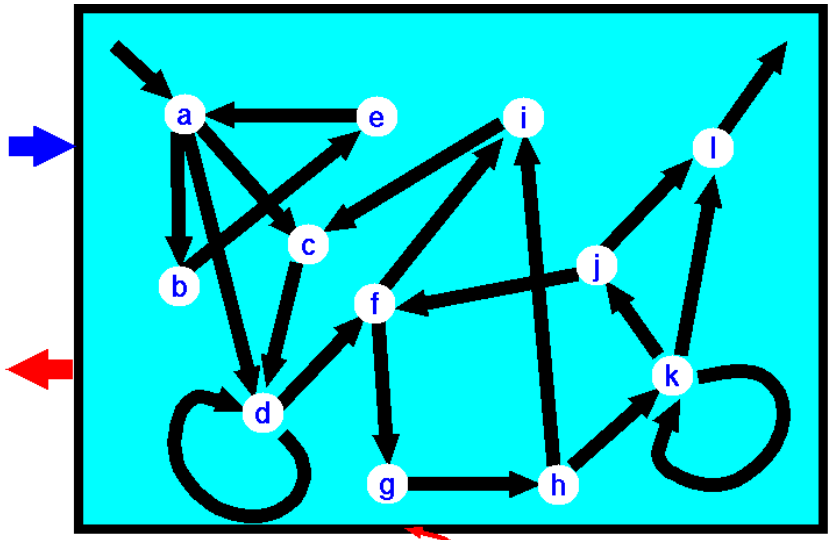
What we have learnt about virtual machines (e.g. programs running on computers), provides new ways of thinking about this – especially AI virtual machines



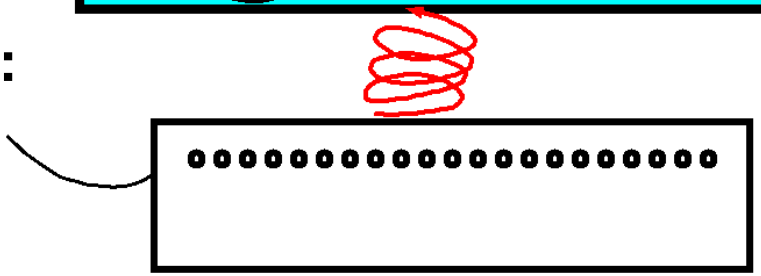
Many people have explored this analogy, but when philosophers use over-simplified ideas about virtual machines they produce over-simplified theories.

How some philosophers think of virtual machines: Finite State Machines (FSMs) (e.g. Ned Block once)

Virtual machine:



Implementation relation:



Physical computer:

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

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

A richer model: Multiple interacting FSMs

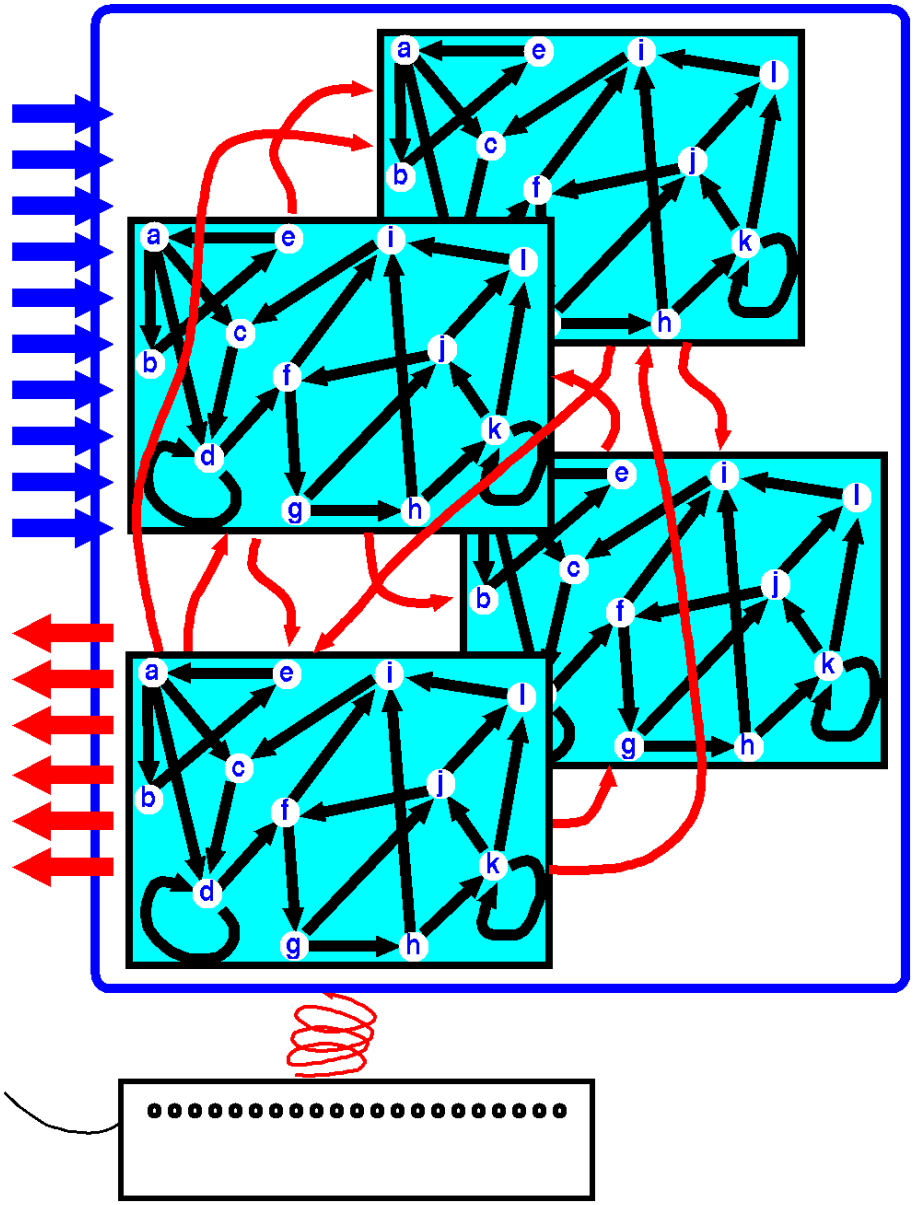
This is a more realistic picture of what goes on in current computers:

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

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

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

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



First, a possible objection

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

This ignores the important role of memory mechanisms in computers.

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

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

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

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

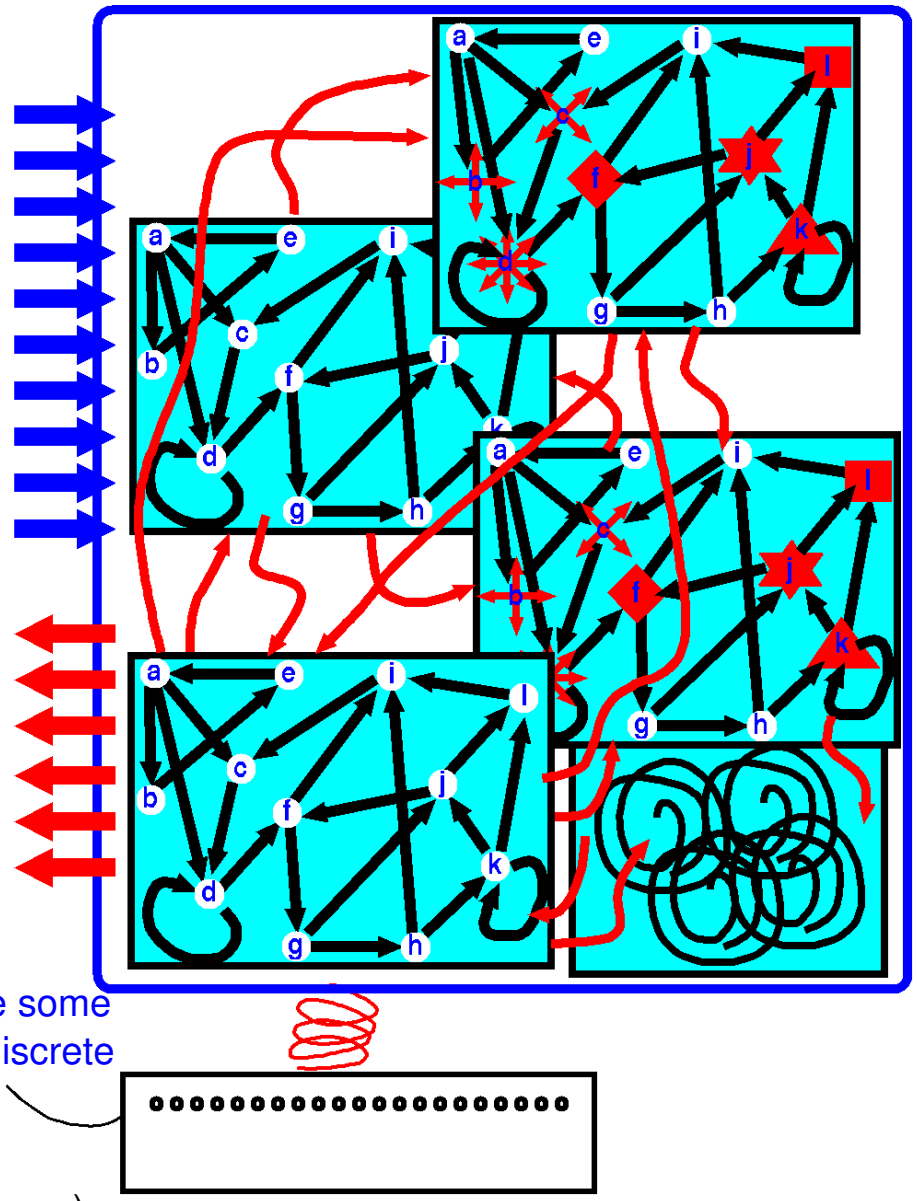
(A technical point: software interrupt handlers connected to physical devices that are constantly on, e.g. keyboard and mouse interfaces, video cameras, etc., mean that some processes are constantly “watching” the environment even when they don’t have control of the CPU.)

A more general model

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

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

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



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

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

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

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

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

Philosophers just tend to ignore this when discussing supervenience,

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

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

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

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

Could such virtual machines run on brains?

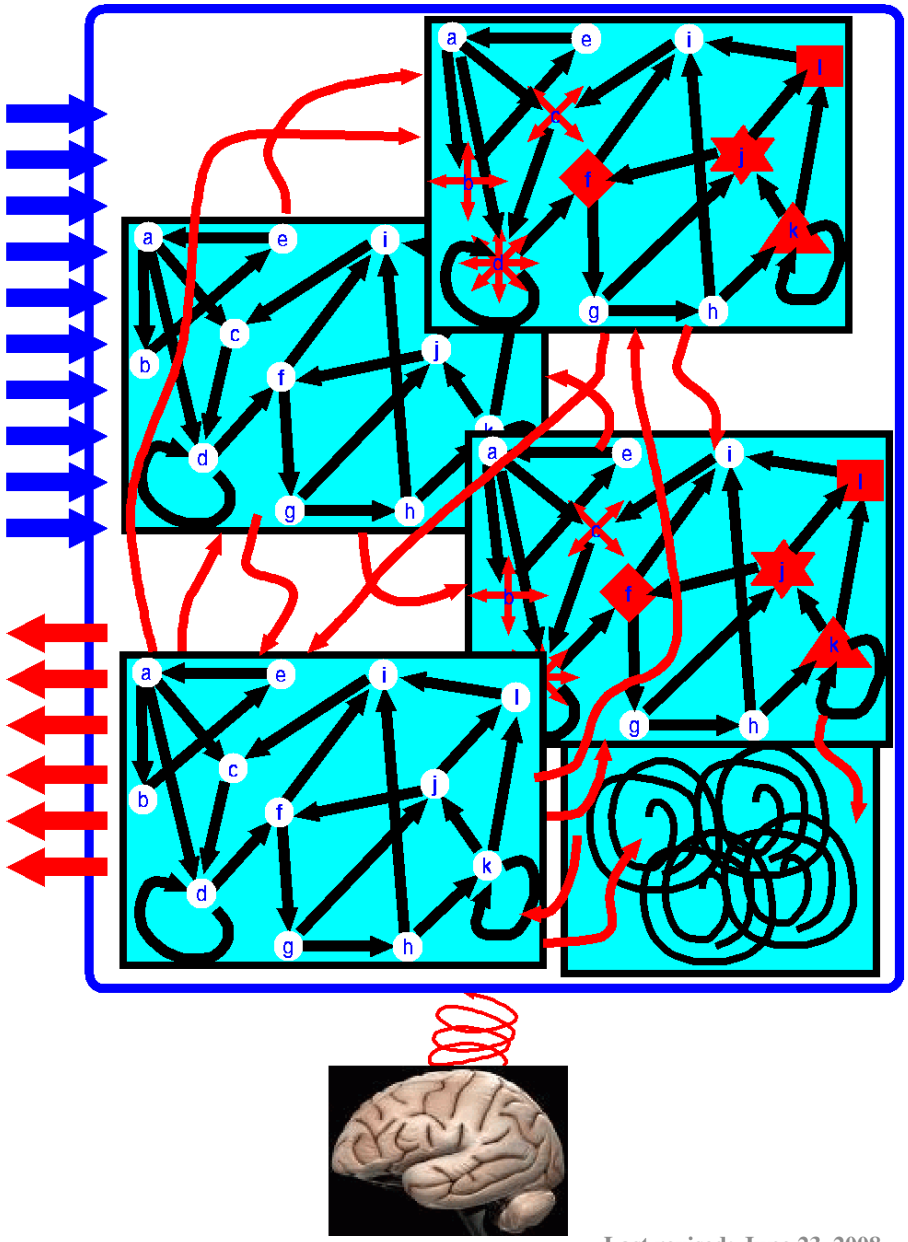
We know that it can be very hard to control directly all the low level physical processes going on in a complex machine: so it can often be useful to introduce a virtual machine that is much simpler and easier to control.

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

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

Questions:

- How would the genome specify construction of virtual machines?
- Could there be things in DNA, or in epigenetic control systems, that we have not yet dreamed of?



Self-monitoring and virtual machines

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

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

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

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

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

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

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

THAT CAN LEAD TO DISASTERS, BUT MOSTLY DOES NOT.

Robot philosophers

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

Intelligent robots will start thinking about these issues.

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

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

BUT WE HAD BETTER DEVELOP GOOD PHILOSOPHICAL THEORIES FIRST!

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

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

For more on robots becoming philosophers of different sorts see

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

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

AI Theorists make philosophical mistakes

A well known "hypothesis" formulated by two leading AI theorists, Allen Newell and Herbert Simon, The Physical Symbol System Hypothesis, states that:

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

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

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

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

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

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

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

A PICTURE OF YOUR MIND?

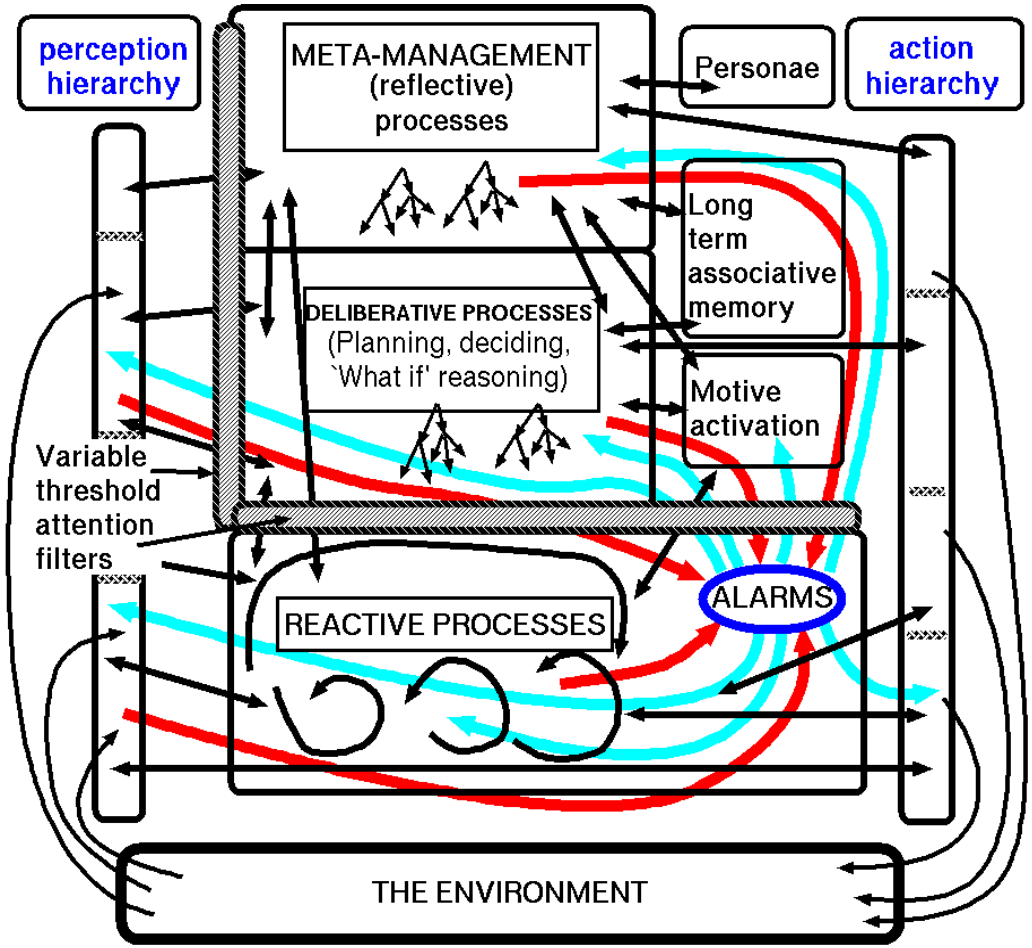
What sort of virtual machine runs on your brain? Here's a crude picture: The H-CogAff architecture.

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

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

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

So far only small parts of this have been implemented.



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

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

- It would have a lot of innate or highly trained reactive behaviours.
- It might have to grow new competences, extending its architecture, as a result of interacting with the environment

  As partially explained in some joint papers with Jackie Chappell, e.g.:
  `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609`
  Natural and artificial meta-configured altricial information-processing systems
- It would be able to do some planning, explaining, predicting, hypothesising, designing, story telling, using its deliberative mechanisms.
- Its metamanagement methods examining and controlling the robot's own high level virtual machine, as well as perhaps thinking about and communicating with others, would probably under some circumstances start doing philosophical speculation about the nature of its own mind.
- The result will probably be a lot of deep philosophical confusion.
- Unless we can teach it to be a good philosopher.
- For a start, we could ask it to study and analyse these slides and evaluate them as presenting a theory about how the robot works.
- Maybe some of them will come up with much better philosophical theories about minds and bodies than any human philosophers have done.

THANK YOU!

For a lot more on supervenience and virtual machines see

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

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

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

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

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

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

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

Additional papers and presentations

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

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

See also the URLs on earlier slides, e.g. Slide 2.