Presentation at Symposium on Machine Consciousness
AAAI'07 Fall Symposium, Washington 9-11th Nov 2007

DRAFT - to be revised

Why Some Machines May Need Qualia and How They Can Have Them:

Including a Demanding New Turing Test for Robot Philosophers

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

These slides will be in my 'talks' directory:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/

The position paper for the symposium is available here
http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0705

Related talk to the symposium on Development and Representations is also online
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#devrep
Diversity of Developmental Trajectories in Natural and Artificial Intelligence

Apologies

I apologise

- For slides that are too cluttered: I write my slides so that they can be read by people who did not attend the presentation.

  So please ignore what's on the screen unless I draw attention to something.

NO APOLOGIES
For using linux and latex

We are hiring!

We urgently need to replace someone working on the CoSy robot project
`http://www.cs.bham.ac.uk/research/projects/cosy/PlayMate-start.html`
with the possibility of working for several years on a follow-on project.

The task involves 3-D vision and manipulation. Outstanding, experienced, applicants should write immediately to

Jeremy Wyatt J.L.Wyatt@cs.bham.ac.uk
Aaron Sloman A.Sloman@cs.bham.ac.uk

I am not trying to build a particular system
I am trying to do something different:

Instead of only trying to model particular organisms, or produce machines to solve particular practical problems

some of us should try to understand **the space of possible systems,**

and the tradeoffs between different designs in relation to different requirements – including the requirements of different organisms

For this we need to understand evolutionary and developmental trajectories

Design space(s) Niche space(s) and their relationships

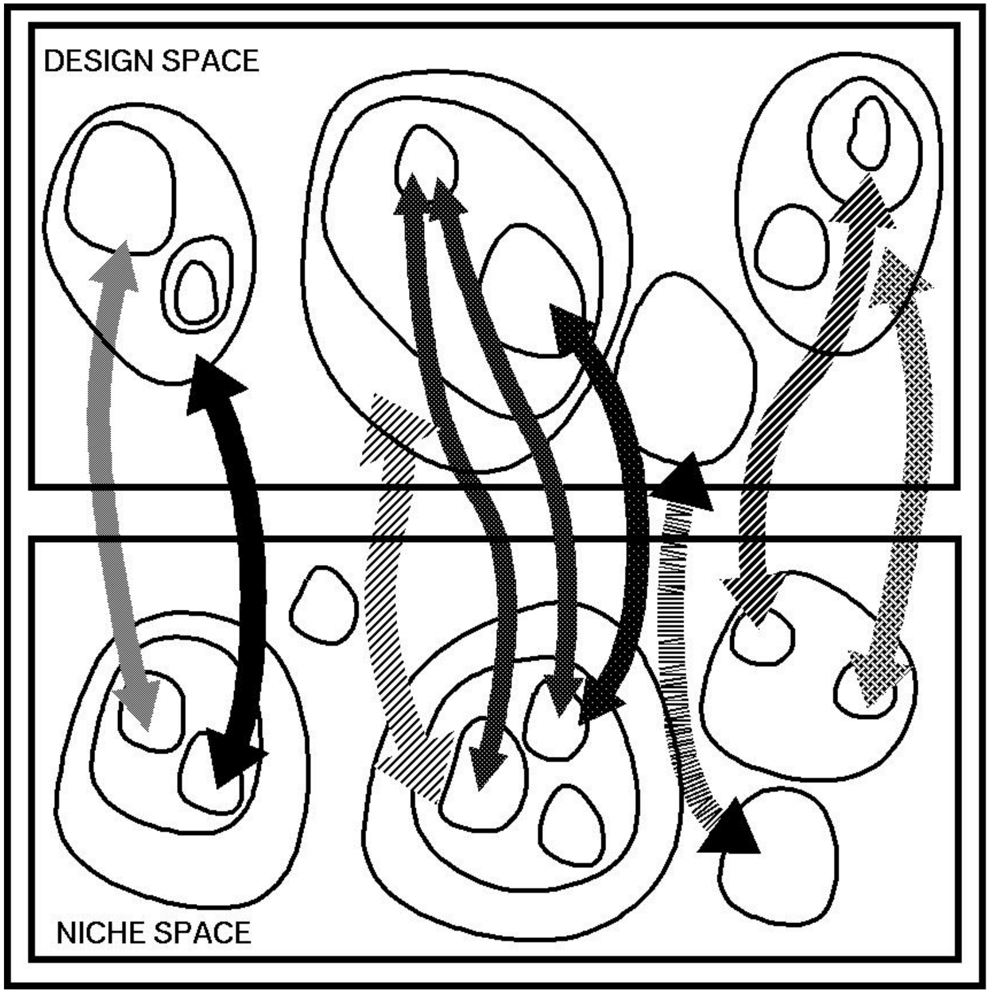
Design space: a space of possible architectures (including mechanisms, formalisms, etc.)

Niche space: a space of possible sets of requirements for whole animals, robots...

There are discontinuities in both design spaces and niche spaces: not all changes are continuous.

Do not expect one fitness function. Instead expect diverse structured fitness relations between designs and niches.

We can also talk about designs and niches for parts of an existing system: e.g. the niche for a digestive subsystem, or a motor control subsystem, or a perceptual subsystem depends in part on what is already in the rest of the machine or animal.



Thus not only species and whole organisms but also subsystems and their designs can co-evolve and co-develop.

Trajectories in both spaces

There are different sorts of trajectories through the two spaces.

i-trajectory: possible for an individual organism or machine, via development, adaptation and learning processes (of many types): egg to chicken, acorn to oak tree, etc.

e-trajectory: possible for a sequence of designs evolving through natural or artificial evolution. Requires multiple re-starts in slightly different locations.

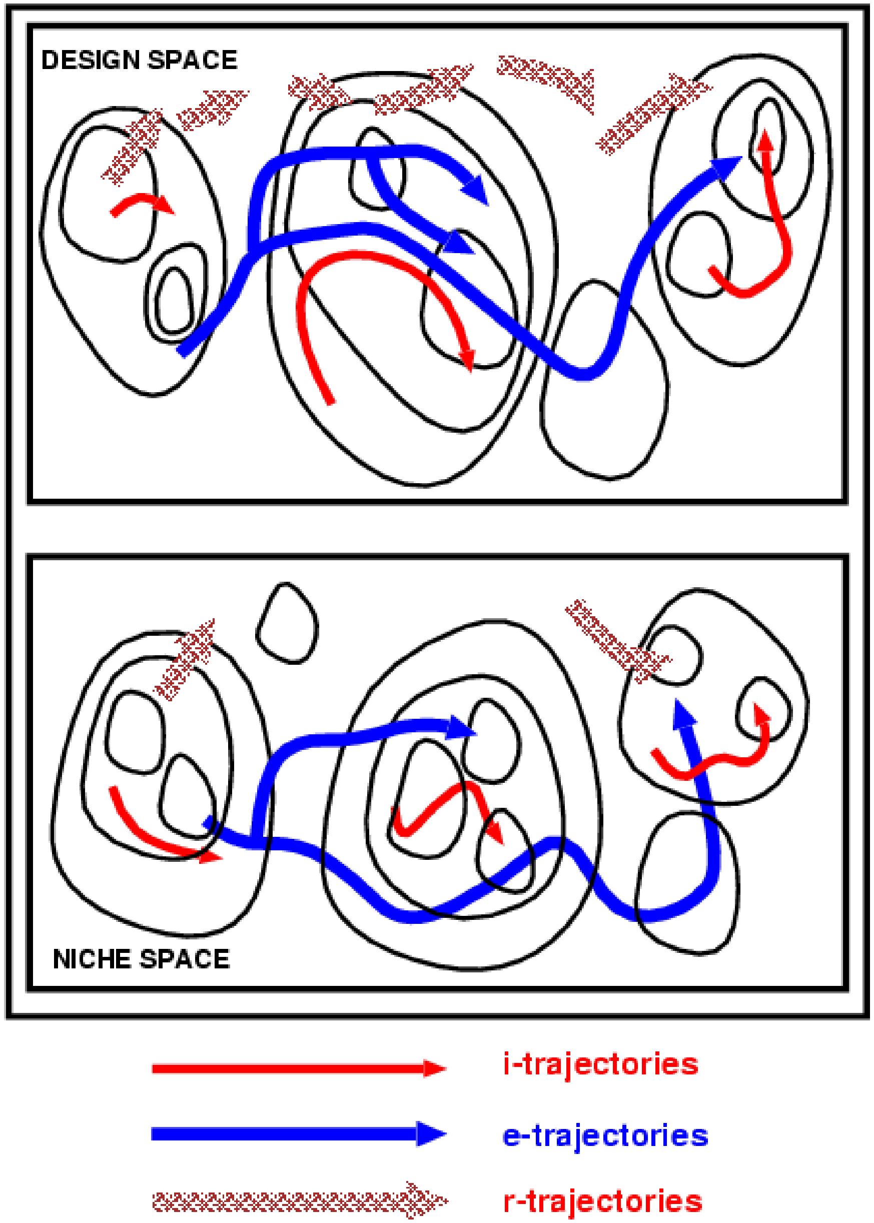
r-trajectory: possible for a system being repaired or built by an external designer whose actions turn non-functioning part-built systems into functioning wholes, or add a new feature: can produce discontinuous trajectories.

---

s-trajectory: possible for social systems with multiple communicating individuals. (Can be viewed as a type of i-trajectory.)

c-trajectory: trajectory made possible by the use of cognitive capabilities of individuals, e.g. mate selection or differential parental caring for young of different capabilities.

All but r-trajectories are constrained by the requirement for “viable” systems at every stage.



Dynamics of Linked trajectories

Motion along a trajectory in design space causes and is caused by motion along a trajectory in niche space

This obviously applies to e-trajectories, and less obviously to i-trajectories

> the niches for an unborn foetus, for a newborn infant, a schoolchild, a parent, a professor, etc. are all different

---

Moreover, an individual can instantiate more than one design, satisfying more than one niche: e.g. switching between being

- protector and provider,

  or
- parent and professor

To cope with development of multi-functional designs we can include *composite niches* in niche space, just as there are composite designs in design space.

Composite niches lead to composite designs and vice versa.

Biological evolution:

The history of the biosphere involves multiple interacting e-trajectories for designs and niches, with many interacting feedback loops.

As more and more complex organisms evolved, their i-trajectories became longer and more diverse.

Much later came s-trajectories and c-trajectories,

See my talk for the Representation and Development symposium for more on the need to study varieties of environments, varieties of organisms, design options, tradeoffs, evolutionary affordances, etc.

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

Where does consciousness come in: Everywhere?

The only kind of consciousness that interests me is the functional consciousness that every biological organism has: consciousness of aspects of its environment

(and in some cases, though not all, consciousness of some aspects of itself).

This amounts to being able to acquire and use information about things.

Many computing systems already have such capabilities, including systems that contain running spelling checkers, virus checkers, operating systems, schedulers, file system managers, device drivers, email systems, etc.

At present no artifacts (or very few) have their own goals, preferences, values, interests, etc. driving what they do.

I am not interested in discussing the ‘hard problem’ in the context of AI.

All “solutions” are bogus, as Stevan Harnad stated yesterday.

We agree that they are, though differ as to why they are.

I am interested in the many varieties of consciousness, awareness, detection, sensing that can occur in organisms and machines and which can produce consequences (not necessarily behavioural consequences).

Beware of AI claims to solve the problem of XXXX

XXXX can be “consciousness’, “emotion”, “free will”, “creativity”, ...

Conceptual confusions make it hard to decide what should go into a machine if it is to be described as ‘conscious’, or as ‘having qualia’ — and researchers tend to choose their own favourite specifications.

AI researcher:

Look: my robot/system has XXXX!

Philosopher:

Your system proves nothing of interest because it does not satisfy the correct definition (= my definition) of XXXX

Read Drew McDermott 1981: “Artificial Intelligence meets natural stupidity”.
In Haugeland *Mind Design*

Just citing the definition or theory of some famous philosopher or psychologist ignores the fact that academics in those fields do not agree on definitions.

Many young AI researchers know only the literature recommended by their supervisors – because they transferred at PhD level from some other discipline and had no time to learn more than the minimum required for completing their thesis.

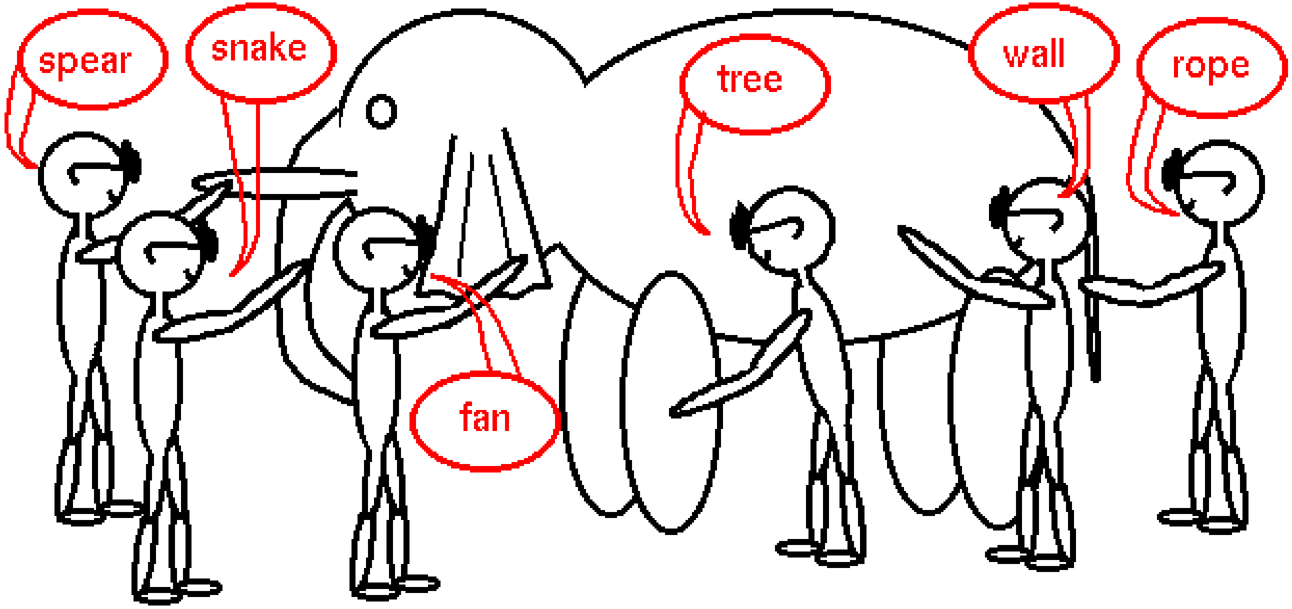
Those who are ignorant of philosophy are doomed to re-invent it – badly.
Apologies to Santayana

We can make progress if we survey design requirements (niches) design options, and the tradeoffs between different designs in different niches.

How to advance knowledge about XXXX

Look at a wide variety of types of organisms, not just humans,
and include infants and genetically impaired, and brain damaged humans,
not just normal adult humans (in your own culture).

Is consciousness an elephant?



Different researchers focus on different features of a very complex system.
But they are unaware of the other features.
Like the proverbial collection of blind men all trying to say what an elephant is:

- One feels the trunk
- One feels a tusk
- One feels an ear
- One feels a leg
- One feels the tail, etc.

Each is correct — about a tiny part of reality.

Read the full poem: by John Godfrey Saxe here:

`http://www.wordinfo.info/words/index/info/view_unit/1`

Ordinary language is OK for ordinary purposes

There's nothing wrong with the ordinary (non-technical) uses of the word "consciousness", and a host of related words and phrases

"attend", "awake", "aware", "consider", "detect", "discern", "enjoy", "experience", "feel", "imagine", "notice", "remember", "see" "self-conscious", "smell", "suffer", "taste", "think", and hundreds more.

They allude to a large number of different kinds of states and processes, serving different biological (or cognitive) functions, and using different sorts of mechanisms, forms of representation, ontologies, and architectures.

- "That fly detected my approaching hand and got away".
- "My dog is aware that I am watching him".
- "While sleep-walking Fred noticed that the door was open, and he shut it".
- "He has been conscious for a few minutes, but is still a bit dopey"
- "I've been aware for months that you don't like me".
- "He felt very self-conscious coming into the room".
- "In my dream I felt frightened"

(Did I or did I not have consciousness then?)

There's no single "IT" to be modelled

It's just bad philosophy to assume that there's some UNIQUE common "thing" underlying all these states, that everything either has or does not have about which we can ask:

- How did IT evolve?
- Which animals/machines have IT?
- When does a human foetus acquire IT?

In philosophical jargon: "consciousness" is a cluster concept

Compare

Wittgenstein: family resemblance concept

Waismann: open textured concept

Minsky: suitcase concept. dumbbell concept

The attempt to identify "IT" always ends up with a "private ostensive definition"

THIS is what I am talking about

Compare: THIS point of space is what I am talking about.

Pointing cannot identify the referent of a multiply-relational concept.

What Turing did NOT do

It is often thought that Turing was proposing a test for intelligence.

He was much too intelligent to do such a silly thing: he said himself that you cannot propose a test for satisfaction of a very ill-defined concept.

He explicitly rejected any possibility of defining “intelligence”, despite frequent claims that that is what he tried to do.

Instead he set up a technological prediction in order to discuss various sorts of objections to his prediction.

In fact his prediction came true: before the end of the century it was possible to fool some high proportion of the general public for a few minutes!

Unfortunately, some AI researchers manage to fool themselves for longer.

Overlap with Owen Holland’s Ideas

To be added later.

I suggest we STOP talking about consciousness

Instead, produce comprehensive sets of detailed requirements for many kinds of functionality — including those found in many different sorts of organisms, and try to understand what sorts of designs can account for them.

If we do that properly every substantive problem of consciousness will be included.

The so-called “hard” problem will remain unsolved because it has been posed in such a way as to make solution impossible.

If we focus on ill-defined global specifications, using words like “consciousness”, “emotion”, “intelligence”, we’ll continue with a morass of inconsistent definitions, goals, and proposed solutions and endless circular debates.

I think Turing understood that well, which is why he explicitly rejected the notion that “intelligence” could be defined, or that there could be a **test** for intelligence.

Show Demos

Trains

Parrot

teach our students to build different sorts of toys to get a feel for architectures, mechanisms, design issues, etc.

But don’t over-interpret them

Sheepdog demo

Feelings semo

We are not merely talking about designs for physical machines

One of the things we have learnt in the last half century is the importance of virtual machines in providing an intermediate level between problems and solutions.

This has powerfully extended our ability to specify, design, debug, modify, and explain complex systems.

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.

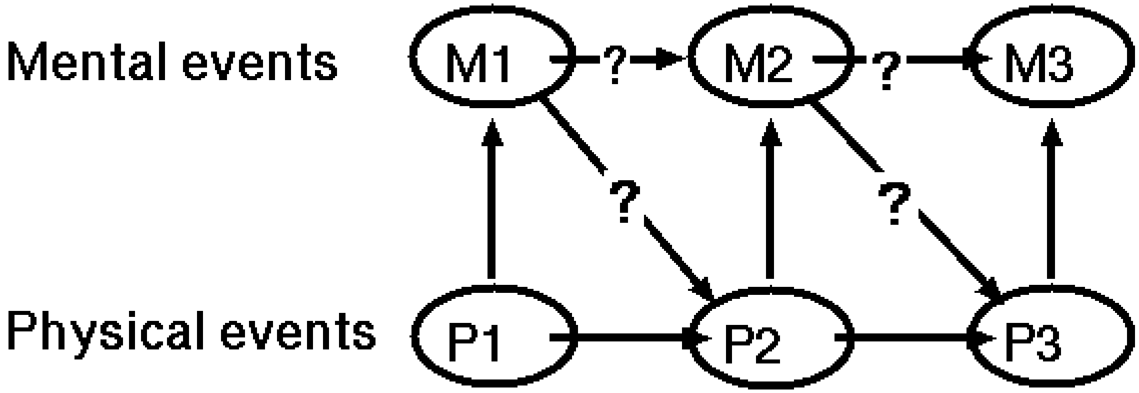
Supervenience and the mind-body relation

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?

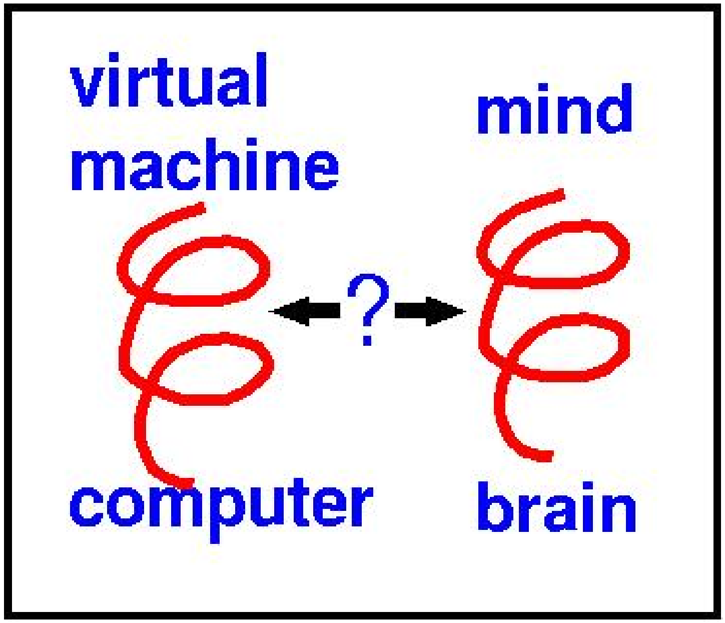


(Think of time going from left to right)

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



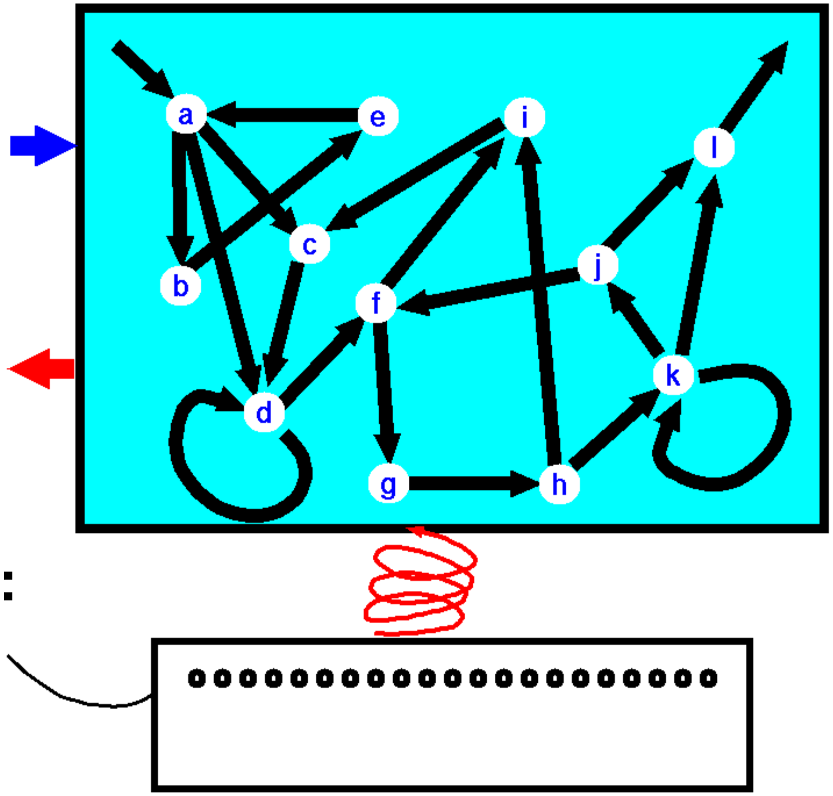
Many people have explored this analogy, but when philosophers use over-simplified ideas about virtual machines they produce over-simplified theories.

How some philosophers think of virtual machines: Finite State Machines (FSMs)

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

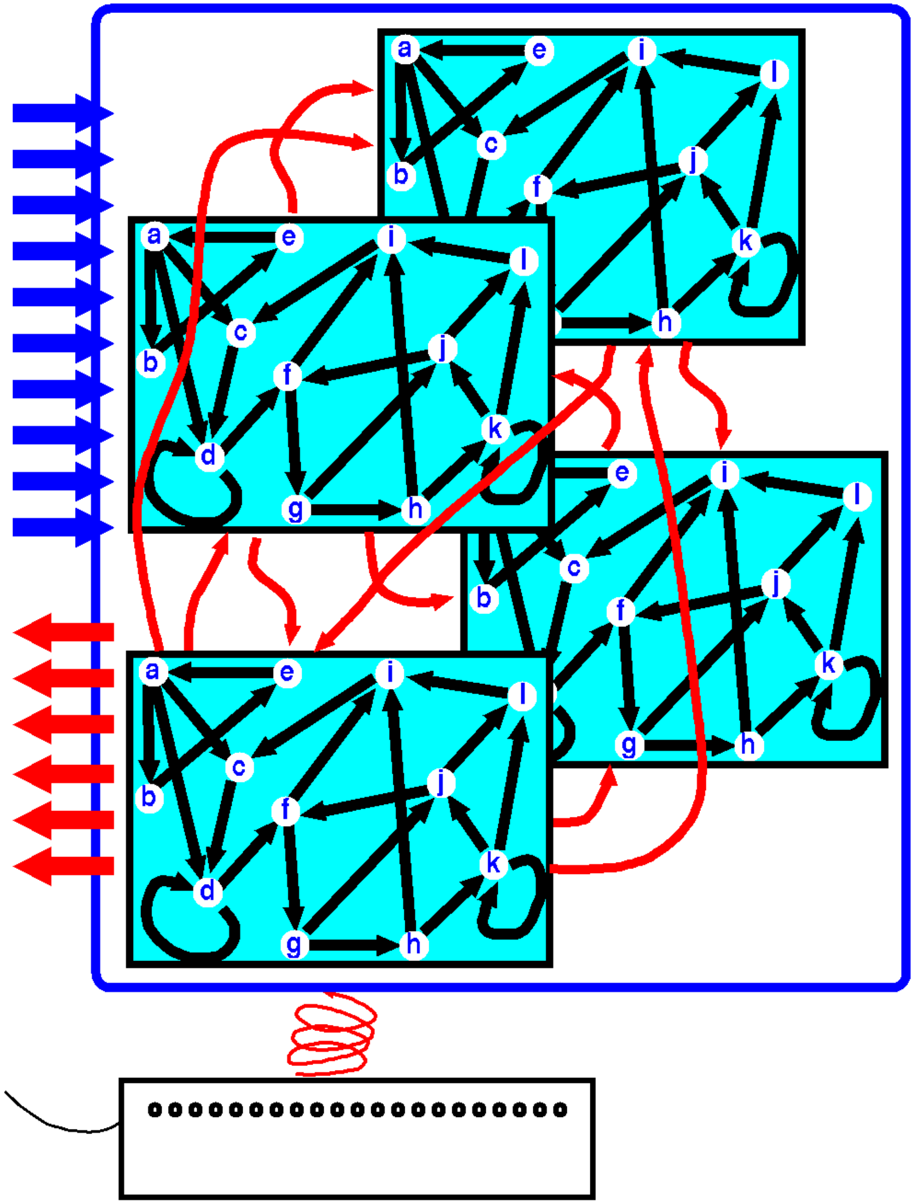
This is a more realistic picture of what goes on in current computers:

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.



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.

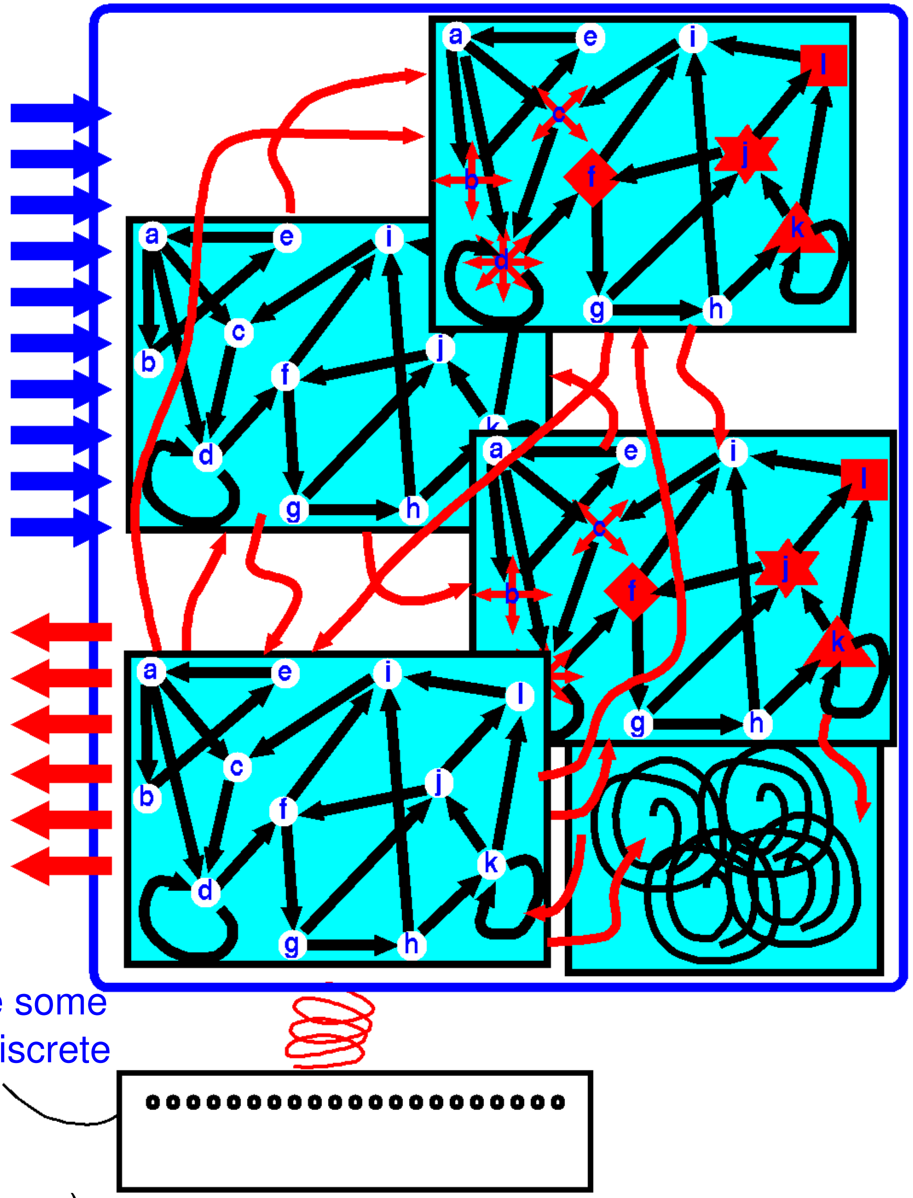
(A technical point: software interrupt handlers connected to physical devices that are constantly on, e.g. keyboard and mouse interfaces, video cameras, etc., mean that some processes are constantly “watching” the environment even when they don’t have control of the CPU.)

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.

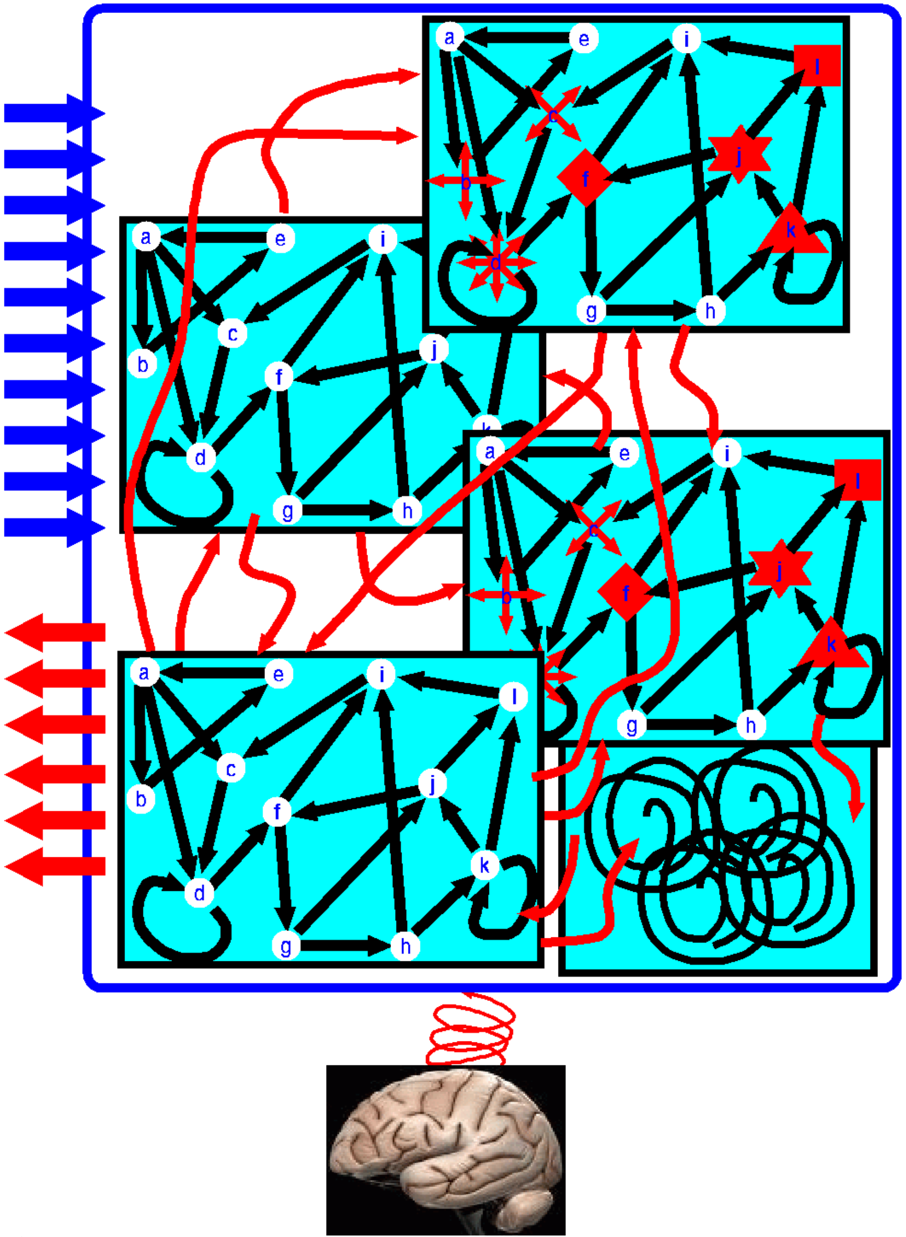
Perhaps evolution discovered the importance of using virtual machines to control very complex systems before we did?

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?



Mind/Body identity explains nothing

One way of trying to avoid the problem (demonstrated yesterday) is to attempt to abolish the mind-body distinction.

But the virtual machine/Physical machine distinction is important for designing and understanding complex systems.

There are too many differences for claims of identity to be useful.

In particular – identity is a symmetric relation, whereas implementation is not.

Tononi's virtual machines

Yesterday Giulio Tononi gave a splendid presentation on families of virtual machines that differ according to

- how integrated they are,

  and
- how much differentiation they can represent
- showing some very interesting tradeoffs.

He related their features to some requirements derived from facts about humans.

This is a very useful contribution that needs to be absorbed into a more comprehensive research programme.

We need to investigate more complete specifications of requirements and the tradeoffs between different design options, instead of just focusing on some particular requirement and toy models meeting those requirements – as many AI systems do.

In particular we need to extend that work to include control mechanisms that turn various subsystems on and off according to changing internal needs and external opportunities and constraints.

We also need to explain how such systems can grow themselves.

Compare the CogAff framework (below)

High level overview of human vision

Vision is a process involving multiple concurrent simulations at different levels of abstraction in (partial) registration with one another and sometimes (when appropriate) in registration with visual sensory data and/or motor signals.

The information is processed in different ways for different purposes, at the same time using different forms of representation.

We don't just see what exists (including **processes**), but also many possibilities for change and constraints on change.

Hence the importance of our ability to see empty space: many things can happen in empty space.

(Compare Sloman and Chrisley, JCS 2003)

But vision is not just something instantaneous: an animal needs to integrate what it sees across saccades and various movements.

Each instantaneous state of the retina (or V1) is just an sample of the ongoing processes in the environment.

The samples need to contribute to an enduring understanding of a constantly changing environment – with some unchanging features.

Compare the theory of Arnold Trehub (2001)

`http://www.people.umass.edu/trehub/`

A new kind of dynamical system (From CompMod07 Talk)

We seem to need a class of dynamical systems

- composed of multiple smaller multi-stable dynamical systems, changing concurrently
- that can be turned on and off as needed,
- some with only discrete attractors, others capable of changing continuously,
- many of them inert or disabled most of the time, but capable of being turned on or off (sometimes very quickly)
- each capable of being influenced by other sub-systems or sensory input or current goals, i.e. turned on, then kicked into new states or processes by bottom up, top down or sideways influences.
- constrained in parallel by many other multi-stable sub-systems
- with mechanisms for interpreting configurations of subsystem-states as representing scene structures and affordances, and changing configurations as representing processes
- where the whole system is capable of growing new sub-systems,
  - permanent
    or
    temporary,
  - short-term (for the current environment)
    or
    long term (when learning to perceive new things).

Competences need to be assembled in an architecture

How can we put everything together?

We need to adopt the design stance and make significant use of present and future concepts from information engineering and science.

That will reveal a logical topography underlying the logical geography of concepts currently in use, pointing at the possibility of new deeper conceptualisations, as happened to ordinary concepts of kinds of stuff, following discoveries about the architecture of matter.

See

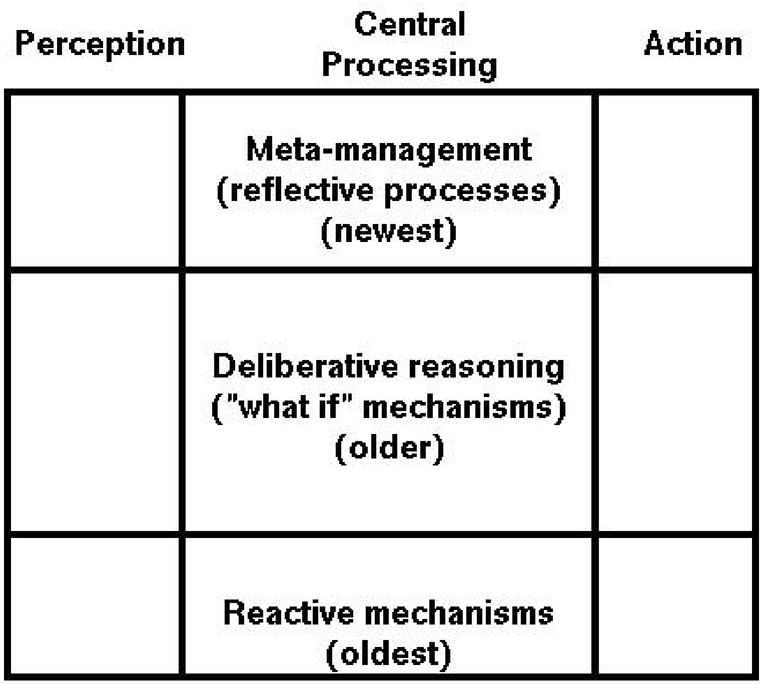
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html`

The CogAff framework – a start

The Birmingham Cognition and Affect project has produced a draft high level classification of types of mechanisms, requiring many further subdivisions (the CogAff schema):

Requirements for subsystems can refer to

- Types of information used (ontology used: processes, events, objects, relations, causes, functions, affordances, meta-semantic....)
- Forms of representation (continuous, discrete, Fregean, diagrammatic, distributed, dynamical, compiled, interpreted...)
- Uses of information (controlling, modulating, describing, planning, executing, teaching, questioning, instructing, communicating...)
- Types of mechanism (many examples have already been explored – there are probably more to be discovered...).
- Ways of putting things together in an architecture or sub-architecture, dynamically, statically.



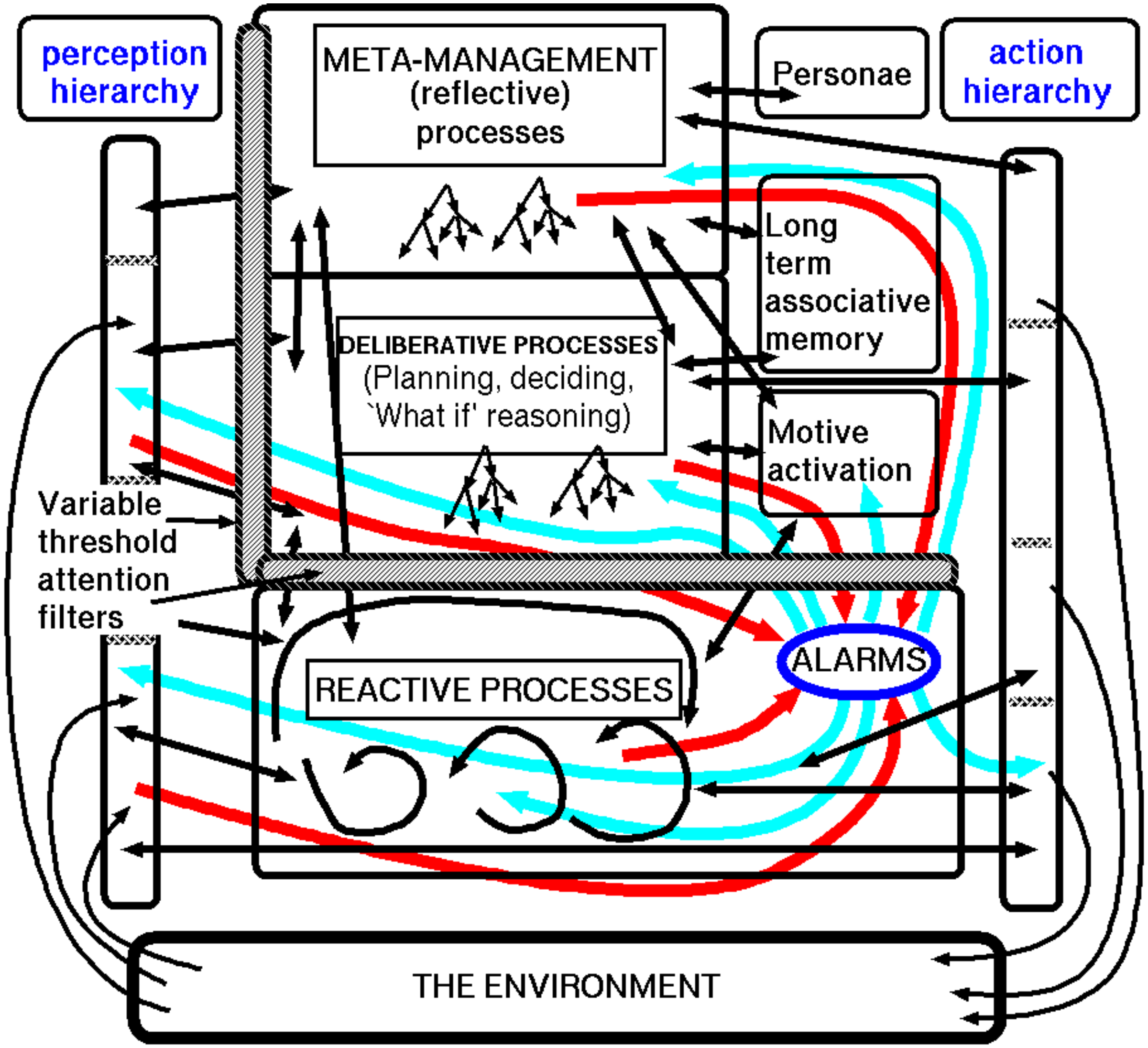
Perception	Central Processing	Action
	Meta-management (reflective processes) (newest)	
	Deliberative reasoning ("what if" mechanisms) (older)	
	Reactive mechanisms (oldest)	

In different organisms or machines, boxes contain different mechanisms, with different functions and connectivity, with or without various forms of learning.

In some, the architecture grows itself after birth.

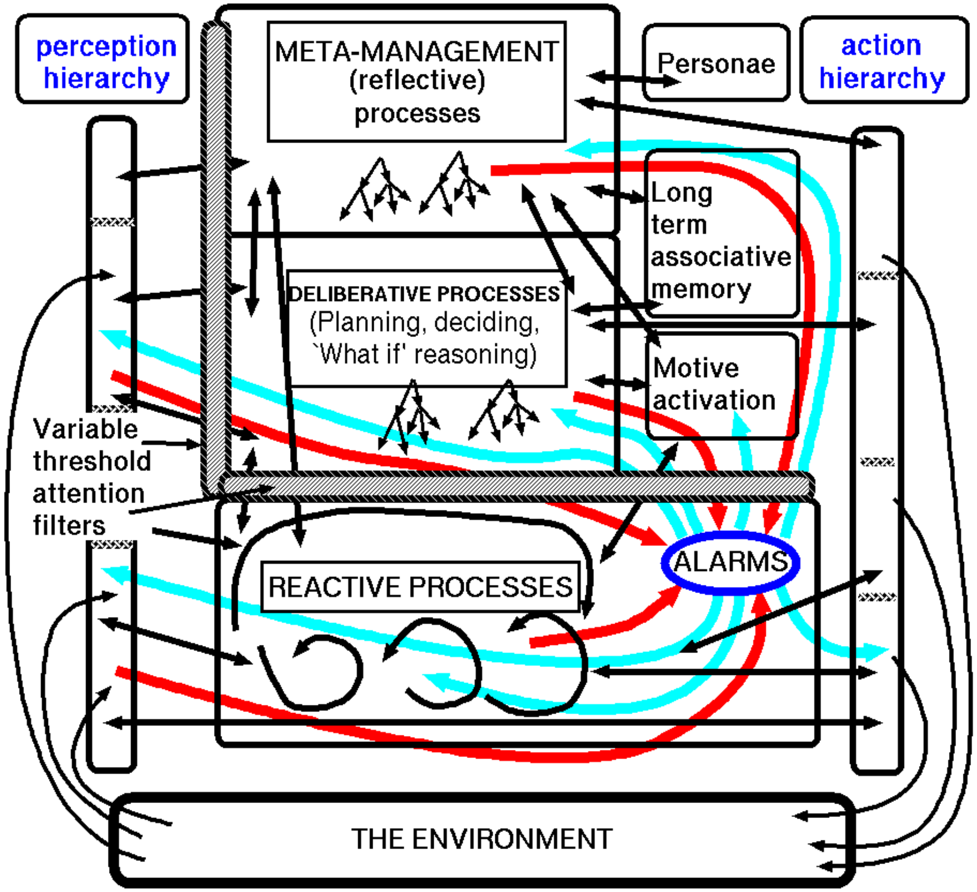
A special case of the schema: H-CogAff

H-CogAff specifies a sub-class of human-like architectures within the generic “CogAff” schema. (“H” stands for “Human”)


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

More about H-CogAff

This is a sketchy indication of some of the required subsystems and how they are connected. Note the implication that both vision and action subsystems have several different (concurrently active) layers of functionality related to the different central layers/mechanisms.


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

Where could this come from?

Different historical trajectories for different layers:

- evolutionary,
  precocial competences from the genome
- developmental,
  altricial competences and architectures built while interacting with the environment
- adaptive changes, (small adjustments)
- skills compiled through repetition
- social learning, including changing personae...

Much work remains to be done.

Kantian causal understanding and reasoning probably cannot occur in the reactive layers. Why not? Different variants may occur in deliberative and metamanagement layers.

For more details, see the presentations on architectures here: http://www.cs.bham.ac.uk/research/cogaff/talks/

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.

Why use virtual machines for control?

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: INSTEAD IT ADDS TO THE POWER OF THE SYSTEM IN MANY CONTEXTS

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`

But we actually agree on many details.

AI Theorists make philosophical mistakes

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

> A physical symbol system has the necessary and sufficient means for intelligent action.
>
> 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 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. (PISS not PSS?)

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

- It would have a lot of innate or highly trained reactive behaviours.
- It would be able to do some planning, explaining, predicting, hypothesising, designing, story telling, using its deliberative mechanisms.
- It may develop an ontology for describing its own internal states and processes (e.g. sensory states)
- 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.

Developing internal languages and ontologies

Such a robot may develop an ontology for describing its own internal states and processes (e.g. sensory states)

If done using a self-organising neural network it might end up with a set of concepts whose semantics inherently use causal indexicality which would make them incommunicable (ineffable).

See Sloman and Chrisley JCS 2003

If it noticed what it was doing it might develop a theory of qualia as some philosophers have done, and rediscover “the hard problem”.

Another robot built to the same design, with components included only because they meet biological or engineering, might be more impressed by different facts about itself, e.g. that it is implemented in a physical system.

The two robots could end up unable to agree on whether qualia were real or whether their existence could be explained by science, or whether they were capable of having causal functions.

Beyond Turing Tests

The Bifurcation requirement (Bi-,Tri-,...)

Most people who propose targets for AI or tests for whether machines satisfy their goals forget the diversity of humans: a diversity that seems to emerge from a common generic design shared by the species.

That diversity includes disagreements about what human minds are, including disagreements about consciousness, free will, the nature of emotions, etc.

A new requirement for scientific AI is to demonstrate how such diversity of philosophical theories can emerge in intelligent machines with a common design meeting only biological requirements

An adequate theory should provide the basis for designs that can start off similar to young humans and under the influence of different cultures and educational regimes are capable of growing up into adults with as much diversity of tastes, interests and strongly held theories as we find in humans.

In particular, for every major philosophical dispute our robot design should explain how every sort of disputant can arise naturally.

Is this a new requirement for adequacy of AI designs?

Compare the requirement of being able to explain results of brain damage, genetic brain malfunction, etc.

THANK YOU!

For the importance of virtual machines and supervenience see

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

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) ?

My presentation to the Development and Representations symposium

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

Additional papers and presentations

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