Evolution of mind as a feat of computer systems engineering

Lessons from decades of development of virtual machinery, including self-monitoring virtual machinery

Aaron Sloman

School of Computer Science, University of Birmingham, UK

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

These slides will be added to my 'talks' directory:

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#sps11

The full workshop paper is here:

http://www.cs.bham.ac.uk/research/projects/cogaff/11.html#1103

Topics – depending on time

- Demos and videos providing examples to be referred back to during the presentation. Toy sheepdog and Toy emotional agents
- Summarise (inadequately) about 60 years of development in computer systems engineering, including virtual machines.
- Explain significance for systems engineering
- Summarise Darwin's problem and offer a VM-based solution
- Relevance for philosophy
  - Metaphysics (what exists)
  - Varieties of functionalism
  - Causation
  - Mind/brain relations
  - Why qualia must exist in certain sorts of machinery
- The need for philosophers to be better educated, so that they can contribute.
- Some remaining hard problems (if there's time).
  - About the space of possible machines and the role of information
  - About what might have happened in biological evolution
  - About epigenetic mechanisms
  - About how brains and minds actually work

Examples: Videos and simple demos

Toy emotional agents, as in

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

Using the SimAgent toolkit

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

See additional online (non-interactive) videos here

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

Video of Betty, the hook-making New Caledonian crow, illustrating some of what needs to be explained in biological information processing.

See movies of Betty here: `http://users.ox.ac.uk/~kgroup/tools/movies.shtml`
The Behavioural Ecology Research Group, University of Oxford

It seems very clear that Betty knows what she intends to achieve and how to achieve it.
Much of the behaviour is directed and purposeful, with no sign of random movements with associative learning.

Some videos of pre-verbal humans showing various kinds of understanding and confusion:

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

The observable behaviours of young children and many animals indicate the existence of sophisticated information processing mechanisms (very hard to replicate in robots).

This paper argues that the control mechanisms are likely to make use of virtual machines providing intermediate levels of monitoring and control.

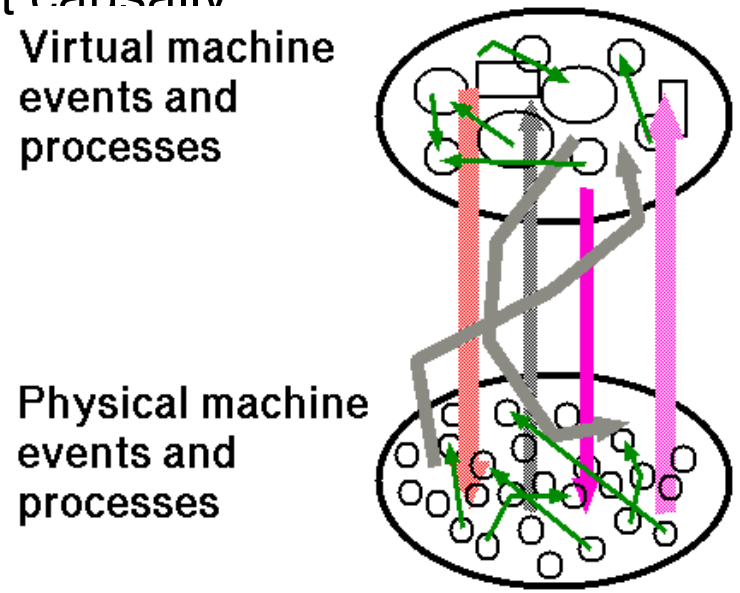
KEY IDEA: What we've learnt to build since about 1950

A web of running virtual machines (VMs) (including VM sub-machines), implemented on physical machines.

Where the VMs and VM components interact causally:

- with one another
- with their physical substrate
- with external entities – physical and virtual
- often using several "layers" of virtual machinery.

Causal influences (discrete and continuous) go upwards, downwards, sideways, and across boundaries – through sensors, effectors, network interfaces, device interfaces...



All this builds on decades of highly creative advances in various sciences, various technologies, mathematics, industrial processes, using computers to help design and build computers...

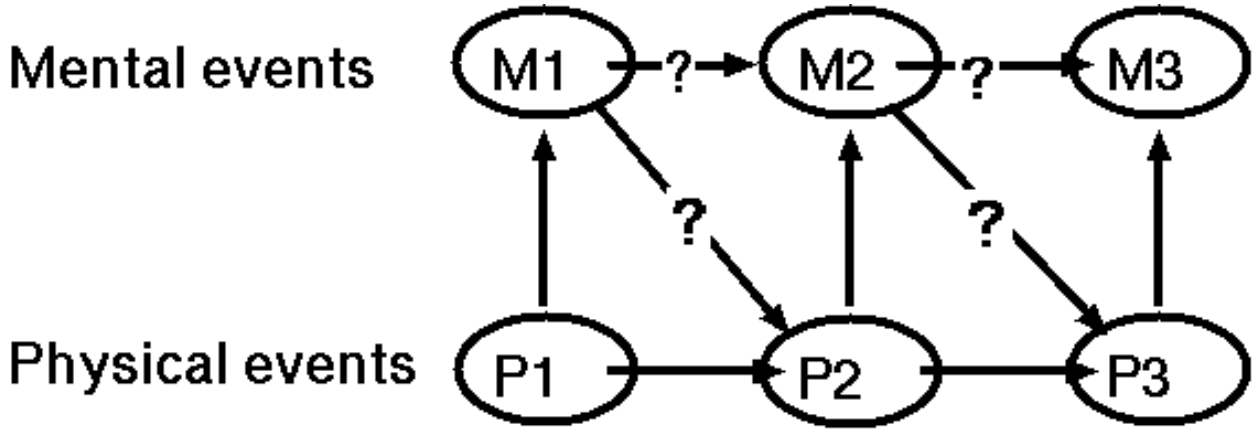
If human engineers can achieve all this in less than a century perhaps biological evolution and development was also able to achieve this – and more – over millions of years?

Compare superb biological mechanical engineering

The wrong model

This is how some people think about relationships between physical events and mental events.

The events are assumed to be discrete and atomic (like transitions in a finite state machine – with one-to-one correspondences):



The causal arrows labelled "?" are thought by many to be dubious.

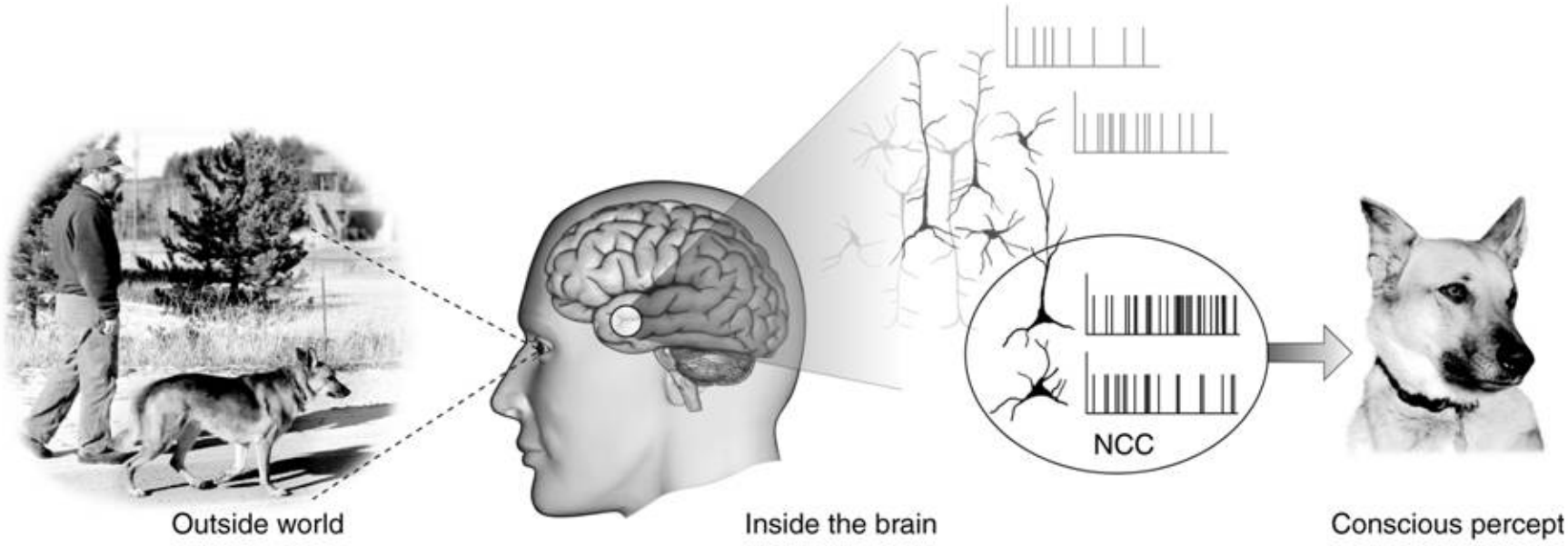
That's one of many wrong models – wrong in ways that will become clearer (I hope).

Philosophy needs to learn from advances in science and technology – not for the first time.

Another wrong model

by Mormann & Koch (From Scholarpedia)

This well known figure summarises a widely-held, but mistaken view of consciousness.



http://www.scholarpedia.org/article/Neural_correlates_of_consciousness

I'll try to explain why processes of the sort suggested

(a) cannot end with images, but require more usable information structures,

(b) cannot end with contents of consciousness, since perception has many purposes including control of behaviour, and that would require the information gained to feed into decision making, and motor control processes (among others).

A common mistake about consciousness

The mistake is to assume that causation goes only one way: P⇒M

We also need to understand how perceptual experiences, pleasures, pains, desires, motives, decisions, intentions, plans, preferences.... can produce physical behaviour?

Epiphenomenalism says they cannot:

It states that mental states are caused but cannot be causes.

The fact that philosophers talk about mental states is evidence that epiphenomenalism is false.

BUT **Mental phenomena are products of biological evolution.** (And other factors)

Mental events and processes have crucial biological roles in control of: actions, energy deployment, goal-selection, plan-formation, plan execution, problem solving.

SO

Theories of mental phenomena must include causal powers of mental events and processes, including

- their ability to cause other mental phenomena
- and their ability to cause physical behaviour

That means we have to explain how “downward causation” is possible.

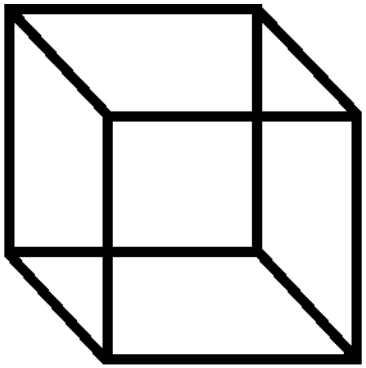
We have learnt a lot about this in the last 50-60 years, but it has mostly gone unnoticed.

NB: This does not imply that all mental processes have behavioural consequences – only the possibility is implied. Not all possibilities are realised.

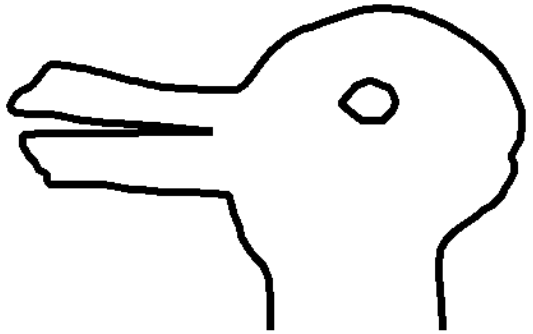
Do some "consciousness science"

Ontologies for conscious contents

Stare at each of these two pictures for a while.



Necker Cube



Duck-rabbit

Each is ambiguous and should flip between (at least) two very different views.

Try to describe exactly what changes when the flip occurs.

What concepts are needed for the different experiences?

In one case geometrical relations and distances change. In the other case geometry is unchanged, but biological functions change. Can a cube be experienced as "looking to left or to right"? If not, why not?

Nothing changes on the screen or in the optical information entering your eyes and brain.

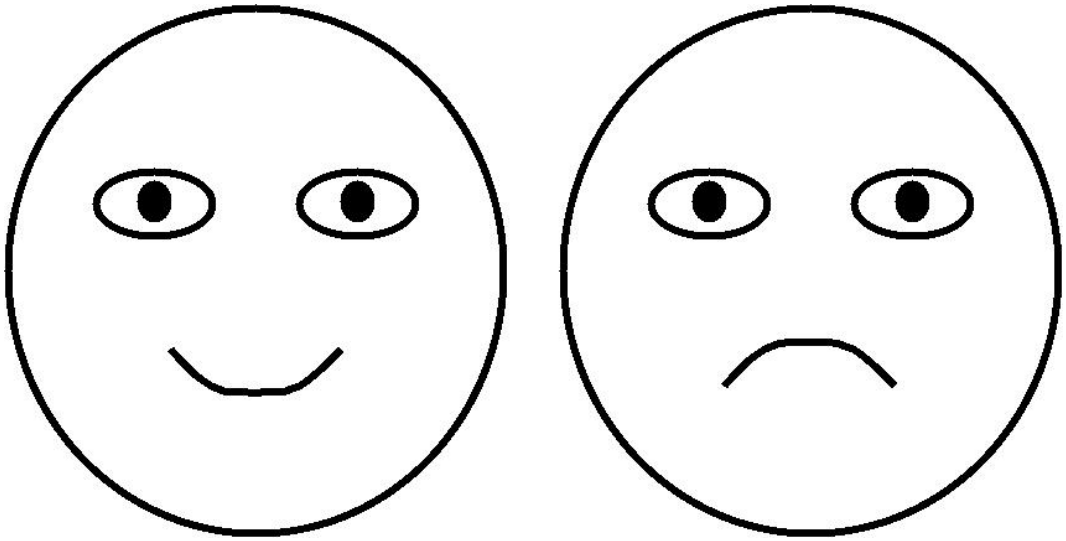
Compare the kind of vocabulary used to describe parts and relationships in the two views of the Necker cube, and in the two views of the "duck-rabbit".

So some contents of consciousness are three dimensional, while others are about entities with functional roles (e.g. eye, mouth, bill), mental states (e.g. looking left), ...

(NOTE: as Morris Sloman reminded me, computers can also be given some information then find two or more interpretations. This is commonplace in computer vision systems.)

Illusions of various sorts also give clues

- Gently press the corner of an eyelid: your experience of the environment (i.e. your current set of qualia) changes, not the environment.
- Move backwards and forwards: what you experience changes, though the objects do not.
- Do the two sets of eyes below look the same?



Many people experience a difference between the eyes in the two faces, despite the eye-images being identical.

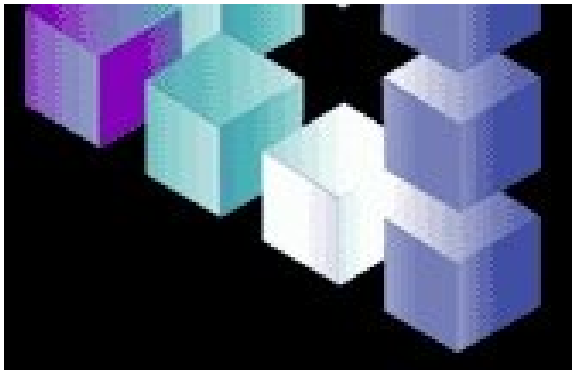
This illustrates the roundabout routes of causation in visual perception.

(Compare Kanizsa's illusory contours.)

Similar effects could occur in a computer vision system that works bottom-up and top-down.

3-D structures – and possible actions

On the right is part of a picture by Swedish artist, Oscar Reutersvärd (1934) which you probably see as a configuration of coloured cubes.



As with the Necker cube you have experiences of both 2-D lines, regions, colours, relationships and also 3-D surfaces, edges, corners, and spatial relationships.

You probably also experience various affordances: places you could touch the surfaces, ways you could grasp and move the various cubes (perhaps some are held floating in place by magnetic fields).

E.g. you can probably imagine swapping two of them, thinking about how you would have to grasp them in the process – e.g. swapping the white one with the cube to the left of it, or the cube to the right of it.

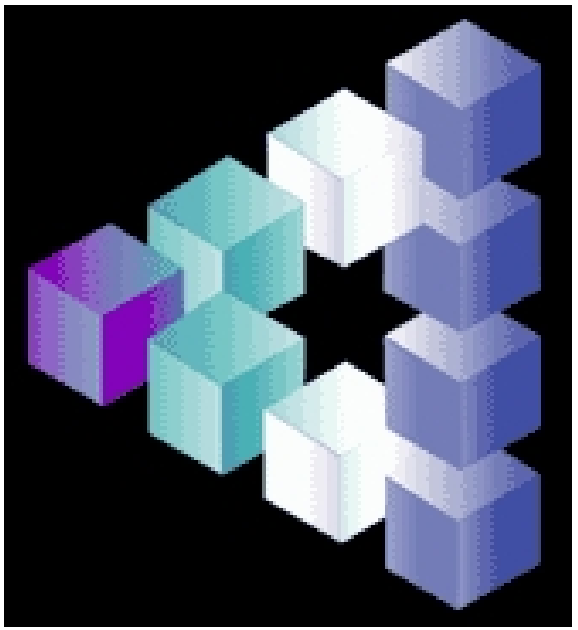
Moreover in an actual 3-D scene looking like that you would be able to use the visual information to perform the actions, not merely think about the possibility of doing so.

This point is closely related to the points made by James Gibson about links between perception and action, and his idea that in animals perception provides information about affordances for the perceiver.

See (Gibson, 1966, 1979) and this slide presentation:
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#gibson`
What's vision for, and how does it work? From Marr (and earlier) to Gibson and Beyond
(Argues that Gibson (and many others) saw only a subset of a complex set of functions of vision.)

2-D and 3-D Qualia: More of Reutersvard

The picture on the right (from which the previous one was extracted) has a richer set of 2-D and 3-D contents.



Again there is a collection of 2-D contents (e.g. a star in the middle), plus experience of 3-D structures, relationships and affordances: with new possibilities for touching surfaces, grasping cubes, moving cubes.

The picture is outside you, as would the cubes be if it were not a picture. But the contents of your experience are in you: a multi-layered set of qualia: 2-D, 3-D and process possibilities.

The scene depicted in the full picture is geometrically impossible, even though the 2-D configuration is possible and exists, on the screen or on paper, if printed: the cubes, however, could not exist like that.

So your qualia can have inconsistent contents!
That's impossible according to some theories of consciousness.

Perceptual qualia are the information structures produced by perceptual systems during the process of analysing and interpreting sensory information.

Qualia can go unnoticed, and then be detected retrospectively. See `http://tinyurl.com/uncsee`

It is likely that many animals are unable to attend to and think about their qualia as (adult) humans can do. What are the biological uses of such capabilities (apart from their use in producing realistic drawings and paintings)?

Conjecture: we can (eventually) explain these phenomena in terms of biological virtual machinery.

The full story is very complex and only a shallow introduction can be given here: There are two main parts:

- An overview of what virtual machines and their contents are
- A suggestion that many experienced contents of visual experiences are structures and processes in virtual machinery.

  In particular, intermediate data-structures in multi-layered perceptual systems of future robots might include both less abstract and more abstract semantic contents, partly in registration with the sensory contents (e.g. retinal images) but also with links to many other information structures, including motive generation, belief formation, decision making, and action control mechanisms.

  There will be several parallel streams of such information forming groupings on different scales at different levels of abstraction, and referring not only to low level sensory contents, but also to things in the environment and relationships, processes, and causal interactions between those things and their parts.

  Some, but not all, portions of those information streams will be accessible to self-monitoring mechanisms.

  Some will be transient – constantly overwritten by new incoming information (e.g. low level sensory buffers, and subsystems concerned with online control of actions – while other perceptual contents will be retained for some time for later uses of various kinds.

We start by summarising some of the features of virtual machinery in computing systems that provide clues.

Summarise about 60 years of development (1)

All modern computer users daily interact with sophisticated virtual machinery running on physical machines (or networks of physical machines) using a web of internal and external causal connections.

Example: What happens when you type a character into a word in a line of text, when using a word-processor?

- Everything to the right of insertion point moves (amount depends on character and font).
- The line may overflow, causing characters to be moved to the beginning of the next line.
- These effects can propagate across several lines.
- They can can cause the current page to overflow, etc. ...
- changing font size, or changing line width or page height can have similar effects – expanding or shrinking number of lines occupied, rearranging the text layout.
- Bit patterns move around in main memory, and contents of hard drives are altered, both in ways that depend partly on other processes.

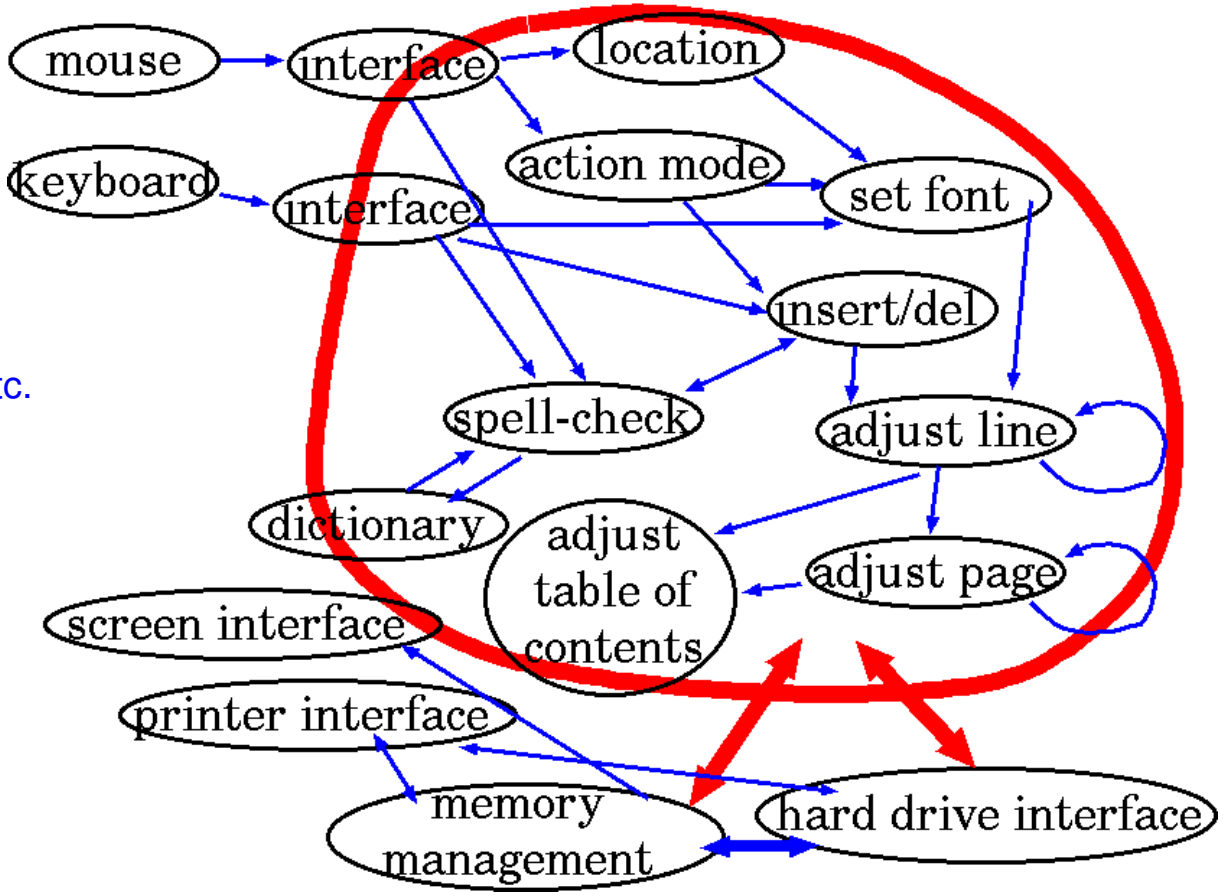
Screen contents, printing information (and possibly audio outputs) are changed.
If the file store is networked, much network activity can be generated.

Some of these are physical events, some VM events.

Part of the text-processor causal web (simplified)

Mouse actions can alter the location and effects of keyboard actions.
Keyboard actions can alter the effects of mouse actions.

Events in processing units,
main memory, connecting buses, etc.
indicated roughly by red boundary.



If the file is sent as an email message far more complex VM and PM events occur, typically involving many machines ... (including virus-checking, for MSWindows users!)

Disclaimer

I have never looked at any code in FrameMaker, Microsoft Word, OpenOffice, or other graphical WIMP (Window, Icon, Menu, Pointing device) word-processors, and don't intend to

though I did work on a much simpler keyboard and screen only editor (Ved, part of Poplog) about 30 years ago.

So what I have written expresses a theory about what's common to a number of virtual machines.

The theory could be tested and found to fit some of them better than others.

Or it might fit none of them.

The example raises interesting issues in the philosophy of science: extra-planetary scientists would presumably regard discovering how human computing systems work as a scientific problem.

That point was recently made in an email message by Harold Thimbleby

However, in the current state of computing technology the sort of virtual machine architecture I have described could be implemented (though the description leaves many gaps).

This illustrates the main point of Chapter 2 of (Sloman, 1978), which argues that discovering what is possible, is a more fundamental function of science than discovering laws and correlations.

Many biological information-processing systems have similar causal webs

There are many forms of information-based control in biology.

The oldest and most pervasive are molecular (i.e. based on chemistry).

They involve networks of interacting collaborative and competitive processes, some regulating/modulating others.

Example networks of concurrent interacting molecular processes are:

- Cell division and reproduction
- Epigenesis (including construction of bodies and brains)
  Cascading, concurrent, waves of development
- Metabolism
- The immune system

My conjecture, below, is that there's much more of the same than has so far been studied, including virtual machine processes implemented in brains.

Summarise about 60 years of development (2)

Changes in lines of text, paragraphs, fonts, font-sizes, sections, chapters, table of contents, spelling errors, grammatical errors, diagrams, captions, etc. are not PM events, they are changes in VM contents.

- The VM is constituted by a rich web of causal relationships between VM components, including many relationships that are conditional on states of other parts of the VM: how a change in X affects Y, can depend on the states or processes in P, Q, R, ...
- Depending on the machine, operating system, and other software/hardware used, inserting a character can (conditionally) cause many more VM events, some producing enduring state changes
  (E.g. has the user's file quota been exceeded? Has the hard drive run out of space?).
- The PM contains only physical devices and their states, including transistors connectors, voltages, currents, electricity being consumed and dissipated as heat, etc.
- Concepts describing VM contents, e.g. "spelling error", "spelling correction", "poor grammar", etc. cannot be defined in terms of the language of physics. Why not?
- However the VM is fully implemented in the PM: replicating the PM (possibly with increases in speed, memory capacity, reliability, etc.) will replicate the VM – no extra ghostly substances are required. **(Sometimes an external context is required – see later.)**
- On current computers there's an intermediate VM layer concerned with operations on **bit-patterns** – partly varying between PMs, partly common.

Significance for systems engineering

Modern computers typically have far more complex virtual machinery running than just a word-processor and associated components.

- Some VMs are application virtual machines performing a specific type of function, while others are platform virtual machines (e.g. components of operating systems) allowing many different additional VMs to be implemented using the platforms.

  All this virtual machinery depends on creative advances of many different kinds in the last 6 decades, by computer scientists, software engineers, device designers, language designers, protocol committees, materials scientists, mechanical engineers, mathematicians, logicians, and many more.

- Some of the advances that support the use of newer VMs would themselves not have been possible without the use of previous VMs that allowed new and diverse internal and external devices to be used without changing high level system designs.

  E.g. the process of reading a file from a hard drive can involve a huge variety of different physical processes as technology changes; but agreed protocols regarding virtual file formats, and file access commands in virtual machines allow new PMs support older VM standards, so old VMs run on new technology and old techniques and systems can go on being used. **(It wasn't always so!)**

  So, when new hard drives are installed using new advanced technology (e.g. higher speeds, greater packing density, etc.), existing programs that need to ask whether the end of a file has been reached don't need to change the VM commands used.

  This is made possible by development of hardware interfaces and software device drivers for new hardware, to ensure required mappings between physical and virtual events and processes.

- So use of VMs enormously simplifies design, development, debugging, self-monitoring, and allowing systems to adapt to new conditions.

Examples of technology supporting running VMs

This is an incomplete fairly arbitrary, possibly misleading, illustrative list, and I make no attempt here to explain what these mean:

- Using addresses rather than physical connections to identify memory locations. (Rapid re-configuration)
- Using memory management systems to control mappings between virtual addresses (in a process) and physical addresses in hardware – including use of garbage collection, paging and other techniques
  – allowing rapid construction and replacement of temporary structures (e.g. for thinking, seeing, etc.),
  – allowing multiple processes to share limited resources in an adaptive (rapidly changing) manner.
- Use of hardware checks, timers, and interrupt handlers to limit what virtual processes can do, enabling violations and errors to be handled cleanly (I.e. no blue screen!)
- Human-readable labels for instructions instead of bit patterns, then increasingly powerful programming languages with more complex types of semantics, supporting more complex applications – e.g.

  named sequences of simple instructions, procedures with recursion and local variables, rule-based systems with rules invoked by pattern matching, languages based on logical inference, use of patterns/templates rather than instructions to specify, recognize, or decompose complex information items, concurrent systems, dynamically varying architectures, hybrid systems, reasoning with simulations, allowing non-determinism, adaptive neural nets and other learning mechanisms, ...
- Use of high level interpreters, compilers, incremental compilers, just-in-time compilers to provide different static and changing relationships between virtual machinery and physical machinery.
- Use of schedulers, accounting systems, distributed operating systems, shared file systems (e.g. using NFS), user and system information shared across networks (NIS/YP) to allow multiple processes, owned by multiple users, with multiple and dynamically changing needs to be securely distributed across multiple networked computers and peripherals.

Sun Microsystems, early 1980s: **"The network is the computer."** (A visionary slogan!)

What is a machine (natural or artificial)?

A machine is a complex enduring entity with parts

(possibly a changing set of parts)

that **interact causally** with other parts, and other "external" things, as they change their properties and relationships.

The internal and external interactions may be

- **discrete** or **continuous,**
- **concurrent** (most machines), or **sequential** (e.g. row of dominoes, a fuse)

NOTE:

Machines, in this general sense, do not have to be artificial, or man-made, or deliberately designed to do what they do.

Physical concepts (PCs) and NPD concepts

"Physical concept" (PC) can be recursively defined thus:

A concept C is a PC if either

(a) C is a primitive concept of physics or

(b) C is explicitly definable in terms of PCs

An explicit definition of a concept specifies conditions for truth or falsity of any proposition using the concept.

Typically fundamental concepts of any science cannot be explicitly defined: they are implicitly defined mainly by their role within the theory.

(Which implies that concept empiricism and symbol-grounding theory are false – as Kant and philosophers of science have pointed out.)

Examples of physical concepts are

mass, length, millisecond, electron, charge, inductance, centre of mass, angular momentum, moment of inertia, energy, momentum....

Not all useful concepts are PCs, e.g.

anger, ignorance, curiosity, need, desire, intention, poverty, crime, democracy, chess, game, win, draw, attempt, economic inflation, international law, newspaper, ...

Non-PC concepts are of different sorts, but this is not the place for a systematic survey: all I need is that there is a distinction between PCs and Non-Physically-Definable concepts, NPD concepts.

Non-physically-definable (NPD) concepts

Certain causally effective states, processes and interactions of some machines

cannot be described using **only** concepts that are definable in terms of concepts of the physical sciences
(E.g. the concepts of physics and chemistry, plus mathematics.)

Many information-processing machines are examples.

E.g. “correcting spelling”, and “attempting to win” (in chess) are NPD concepts.

NB. “Information” is not used here in Shannon’s sense,
but in the sense that includes “reference”, “meaning”, “denotation”, ...
with properties and relations like: **truth, consistency, entailment, contradiction,**
This concept is not **definable** in terms of concepts of the physical sciences.

Though every information-using machine must be implemented (realised) in a physical machine.
For more on the concept of “information” see (Sloman, 2011b)

Concepts of physics can grow and change...

so “non-physically-definable” is a concept that can be revised over time.

Computer Scientists refer to two sorts of VM

We contrast the notion of a PHYSICAL machine with two other notions:

1. a VM which is **an abstract mathematical object** (e.g. the Prolog VM, the Java VM)
2. a VM that is **a running instance of such a mathematical object**, controlling events in a physical machine, e.g. a **running** Prolog or Java VM.
   Running VMs (RVMs) are what this presentation is about.

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

VMs as mathematical objects are much studied in meta-mathematics and theoretical computer science.

They can have complex structures, but are no more causally efficacious than numbers.

The main theorems of computer science, e.g. about computability, complexity, etc. are primarily about mathematical entities

They are applicable to non-mathematical entities with the same structure – but no non-mathematical entity can be proved mathematically to have any particular mathematical properties.

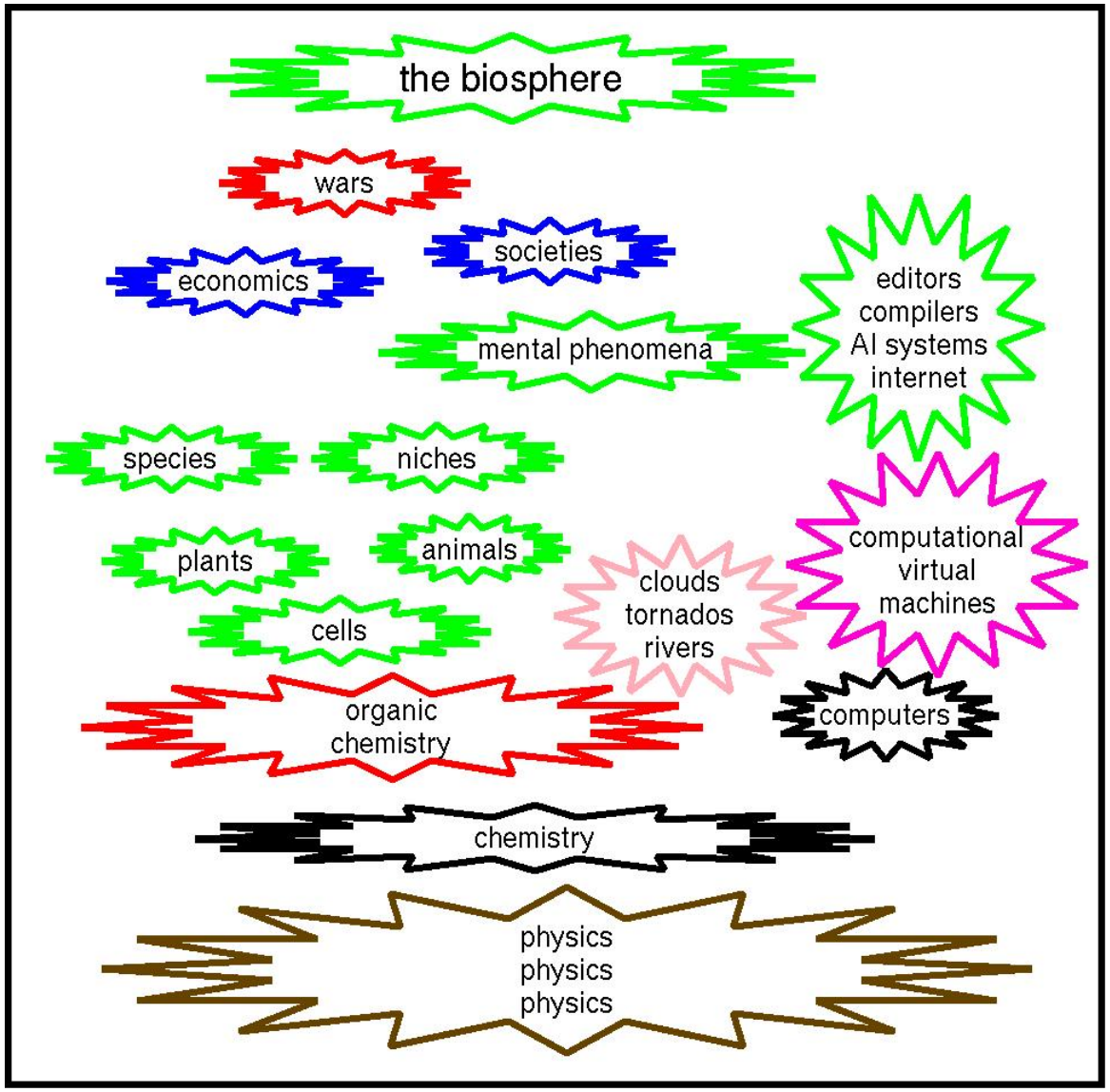
There's more on varieties of running virtual machines (RVMs) in later slides.

Running virtual machines are everywhere

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

E.g. poverty can cause crime.

- All levels are 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?)
- The study of virtual machines in computers is just a special case of more general attempts to describe and explain virtual machines in our world.



NB: Universal Turing Machines are universal only relative to a restricted class of machines.

All RVMs require implementation PMs, but the pre-requisites differ enormously.

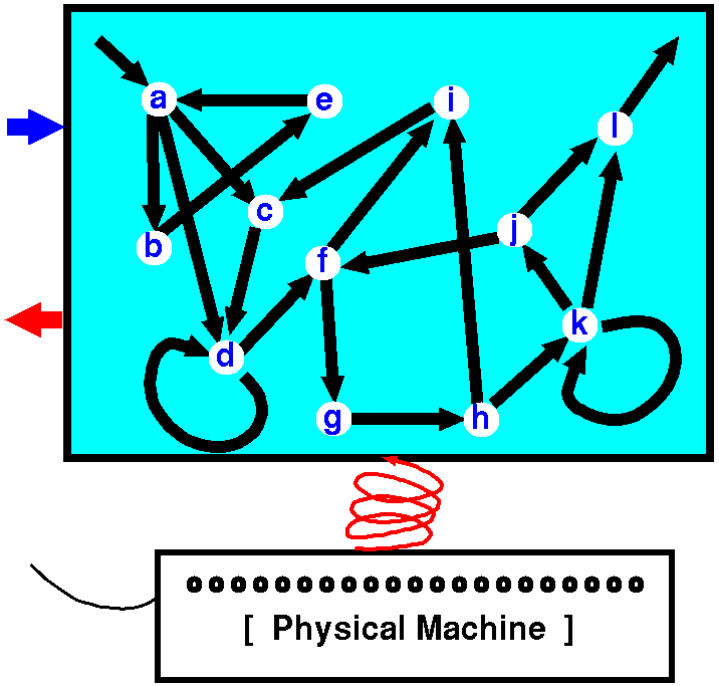
Familiar but inadequate types of VM and functionalism

Standard functionalism invokes Atomic State VMs

"Atomic State Functionalism" assumes everything of interest about a computer reduces to its being a **finite state machine,** following a trajectory through its state space, where each state is physically defined.

The letters label states and the arrows label transitions that can be triggered by inputs (blue arrow). Some transitions can also produce outputs (red arrow).

States in a Turing machine are essentially like this, with the tape replacing the world, and with extremely restricted inputs and outputs.



The trajectory of states in such a VM is closely related to the trajectory of states in the underlying PM.

That's unlike modern computing systems, in which important virtual machine transitions are very different from physical state transitions.

More complex VMs

Typical modern computers are much more complex than that:

- their operations can be decomposed into parallel processes
  e.g. using computer networks, multi-core CPUs and GPUs, and many common-place devices in PCs.
- each with its own state transitions (possibly occurring at different speeds),
- where some of the processes are themselves composed of concurrent sub-systems
- and there are causal interactions between coexisting entities, and processes
- and there are not simple correspondences between the states, state-transitions, enduring entities, causal interactions that are important for the operation of the computer and the underlying physical events and state-transitions.
  (e.g. because of memory management).

We need to learn to see a computer as a complex machine consisting of simpler machines of different sorts (VMs and PMs) interacting with one another and also with the environment.

Some of the simpler machines are PMs
Some are VMs and some are "VM-PM-bridges" (e.g. device drivers).

The relations between PMs and VMs can be far more complex than the simple model (more complex than atomic state virtual machines).

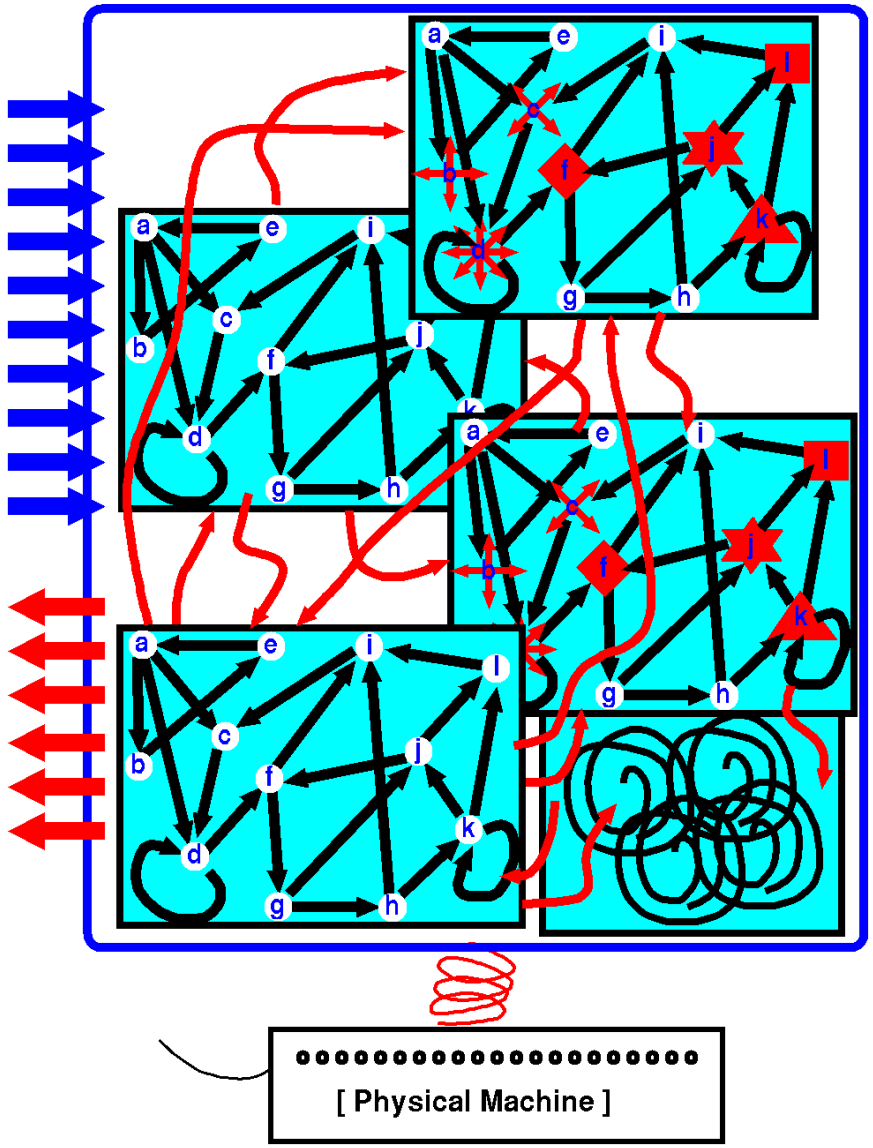
A richer model of a VM

Instead of a single sequential process, a VM can have parts that run in parallel (even asynchronously), interacting with one another and the environment.

Instead of having a fixed set of sub-processes, many computing systems allow new VMs to be constructed dynamically,

- of varying complexity
- some with a short life,
- others going on indefinitely.
- some spawning new sub-processes...
- some discrete, some continuous
- collaborating, competing, or doing unrelated things
- with some transitions probabilistic
- with multiple internal connections (e.g. communication channels)
- a subset connected to external interfaces (possibly sharing input and output devices).

See “The mind as a control system” (Sloman, 1993).



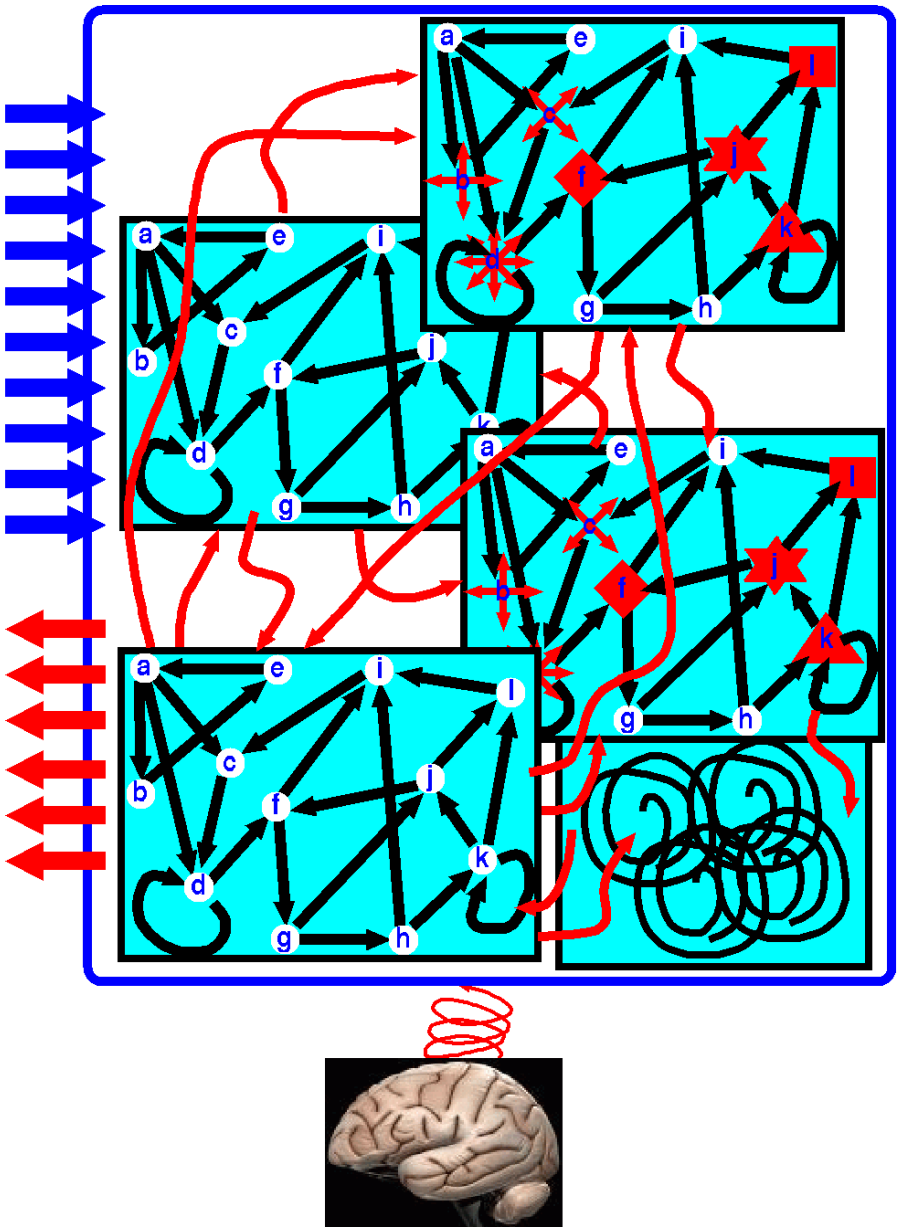
Could such virtual machines run on brains?

It can be very hard to monitor and control all the low level physical processes going on in a complex machine: so it is often useful to introduce virtual machines that are much simpler and easier to control, at the level of their intended functionality.

Perhaps evolution "discovered" the importance of using virtual machines to enable very complex systems to control themselves, long before we did?

So, VMs running on brains (and connected parts of the world) could provide a high level control interface including self-monitoring, and self-modulation.

To any mathematician who uses diagrams, equations, etc. on paper or blackboards, etc., it is obvious that human minds do not run **only** on brains. See (Sloman, 1971).



Current dynamical system models

There are many attempts to replace conventional symbolic AI models and conventional neural net models with models using dynamical systems where internal state transitions are closely coupled through sensory motor signals with the immediate environment.

Such models may be fine for microbes and some insects, but will not do for organisms whose mental processes refer to things in the past, in the future, and to things that could happen but are not happening, to alternative possible action plans only one of which can be selected, and to environmental constraints on possibilities.

The standard dynamical system models developed so far (and also neural net models developed so far) fail to account for human abilities to solve equations, to do logical reasoning, to prove geometrical theorems, to design new forms of machinery, to plan novel buildings, and to construct explanatory theories in physics, chemistry, biology, astronomy, etc.

See (Sloman, 2009c) and
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/multipic-challenge.pdf`

Too often people constructing explanatory theories consider only a small subset of what needs to be explained, and test their theories only on those subsets.

When they choose different subsets they engage in factional battles (e.g. competing for funding, or for students) instead of collaborating in science and philosophy.

"Sense–Think–Act" models are too restrictive

Concurrent interacting VMs remove the need for "Sense–Think–Act" loops.

Many computational modellers and AI researchers assume that an intelligent machine must repeatedly:

- Sense/perceive what is in the environment
- Think about implications and what to do
- Act according to the decisions taken.
- Repeat the process

(Some researchers add more stages, e.g. "reflective" stages, while keeping them sequential.)

This leads to (a) impoverished architectural designs, and (b) a substantial literature on how to optimise the allocation of processor time between the three (or more) stages.

In contrast, in biological brains there is considerable separation of function so that different tasks can be done concurrently (using their own specialised hardware).

Examples include seeing, walking, thinking, and talking concurrently, and integrating use of multiple sensor and effector channels in dynamical systems, e.g. while running, jumping, etc.

Also various kinds of internal self-monitoring, e.g. during speech production and other processes.
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0803`,
Varieties of Meta-cognition in Natural and Artificial Systems

There can be subsystems with sense-think-act loops.

Hard Questions

- How can a genome specify construction of virtual machines?

  (Including layered VMs, and including both platform VMs and application VMs)

  Or partly specify the construction, allowing important aspects of the specification to come from interactions with the environment?

  As in language development.

  (Specifying construction of physical machines is hard enough!)

- What roles can epigenetic environments play?

NB: There is a difference between a genome specifying the final form, and the genome specifying a process of development that can be substantially affected by the environment.
(As suggested in (Karmiloff-Smith, 1992) and (Chappell & Sloman, 2007).)

For example: human abilities to develop linguistic competences seem to depend on features of the human genome not shared with other species.

However, the genome does not specify which language a human should use.

That and many features of the language arise through interaction with the environment.

NOTE:

Language learning is often thought of as a form of data-mining: looking for evidence of vocabulary, syntax, and semantics, in linguistic data.

But there is evidence (e.g. from Nicaraguan deaf children) that it should rather be thought of as a form of collaborative problem solving: creating means to communicate.

Normally a child is in a small minority and the majority have already reached a consensus.

- Could there be things in DNA, or in epigenetic control systems, that we have not yet thought about?

The "Explanatory Gap" (1)

A problem that puzzled Darwin and fired up his critics:

- There's lots of evidence for **evolution of physical forms**.
- There's no similar evidence that human minds could be products of evolution.
- There seems to be no way that physical matter can produce mental processes.
- Even some of his supporters thought natural selection could not produce minds and consciousness
- Wallace – the co-inventor of the theory thought not.
- There seemed to be a serious problem of the relation between mind and matter.

  Much discussed by Darwin's contemporaries – friends and foes.

This is the so-called "explanatory gap"

NOTE

Many young philosophers think the problem of how consciousness can exist in a material world was first identified a decade or two ago, e.g. by Chalmers, or Block, or some other living philosopher.

In fact it goes back centuries. Even the phrase "explanatory gap" goes back to T.H. Huxley, writing in Darwin's time.

For some references and quotations see this talk at SAB2010 (Sloman, 2010b), online here:
`http://www.cs.bham.ac.uk/research/projects/cogaff/10.html#sab`

(There are some examples below.)

The "Explanatory Gap" (2)

New concepts and theories powerful enough to support thinking about how to fill this gap have only been developed in the last half-century or so.

- Until the last few decades, explanatory mechanisms linking physical and mental phenomena were not even conceivable to most scientists: hence the "explanatory gap" of Huxley and others and Chalmers' "Hard problem of consciousness", etc.
- Now, as a result of a great deal of work on hardware, software, firmware, and CS theory, we know how to make things that have some of the features required for working explanatory models of mental processes, with some of the key features of mental processes (including having causal powers, without being physical processes) – but only in very simplified form.
- Most philosophers, psychologists and neuroscientists have ignored or misunderstood this, and so have many AI/Computing/Software researchers.

  I'll give some pointers, but not explain in detail, below.
- There are deep implications for philosophical analyses of causation.

  E.g. **downward** causation from mind-like events and processes to physical events and processes.

Some quotes

Several thinkers in Darwin's time mentioned the problem.

T.H. Huxley

"How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of the Djin when Aladdin rubbed his lamp".
Lessons in Elementary Physiology (1866) (Where exactly?)

G. J. Romanes

"We know by immediate or subjective analysis that consciousness only occurs when a nerve-centre is engaged in such a focusing of vivid or comparatively unusual stimuli as have been described; and when as a preliminary to this focusing or act of discriminative adjustment there arises in the nerve-centre a comparative turmoil of stimuli coursing in more or less unaccustomed directions, and therefore giving rise to a comparative delay in the occurrence of the eventual response. But we are totally in the dark as to the causal connection, if any, between such a state of turmoil in the ganglion and the occurrence of consciousness."(p75)
Mental evolution in animals (1883)
(Quoted by Whittaker in his review in *Mind, 1884*)

There were many others, both before and after Darwin's time, and searching through journals published after *Origin of Species* there is much to be rediscovered about views of the mind/matter problem in the 19th century.

It is much easier to do now that so many old journal issues have been digitised, e.g. *Mind*.

I'll start by helping you understand, at first hand, what all the fuss was about.

MAIN CONJECTURE

- Biological evolution "discovered" many of the problems long before human engineers,
- and produced solutions involving complex VMs whose rationale and operation have not yet been understood,
- using PMs of staggering complexity – supporting extremely complex VMs (and some simpler VMs)
- Some of the drivers of this process may have been
  - The need for concurrent processing of multiple sensory inputs and motor outputs;
  - the need for motive generators and "alarm" systems to be able to operate in parallel with actions based on current motives;
  - the need for animal vision systems to perceive complex changing/interacting 3-D structures in the environment;
  - development of deliberative mechanisms for exploring multiple multi-step possible futures, in order to find a plan or make a prediction;
  - development of self-monitoring, self-modulating control systems.

Often the use of a virtual machine makes kinds of self monitoring and self-modulation possible that would otherwise be impossible, because of the complexity of detail at the physical level.

The separation of levels also facilitates exploration of different designs, by decomposing the problems into simpler problems.

This often involves the use of "platform virtual machines".

Warning from biology:

Don't expect a **sharp** divide between systems using only physical machines and those also using virtual machines: biology provides intermediate cases for most distinctions,

e.g. is a homeostatic control loop a VM?

Neither biology nor engineering needs to respect philosophers' desires for simple classification schemes:

there tend to be many small discontinuities rather than just a few big ones.

But differences across multiple steps can be huge.

Relevance for philosophy

- Metaphysics: what exists
- Epistemology: what concepts are available and what explanatory theories can those concepts be used in?
- Causation: can virtual machines, or their components, really be causes?
- Varieties of functionalism
  atomic state functionalism (ASF)
  virtual machine functionalism (VMF)
- Mind/brain relations
  varieties of supervenience
  Virtual machine supervenience
- Why qualia must exist in certain sorts of machinery

The need for philosophers to be better educated,

So that they can contribute.

See the final section of the conference paper:

`http://www.cs.bham.ac.uk/research/projects/cogaff/11.html#1103`

Some remaining hard problems.

- About the space of possible machines and the role of information
- About what might have happened in biological evolution
- About epigenetic mechanisms
- About how brains and minds actually work
- About whether the required kinds of causal interaction between portions of a virtual machine (e.g. concurrent opposed pressures, such as conflicting desires) can be properly implemented in computer-based systems.

  Current implementations using numerical evaluations of strength or importance to feed into a decision maker do not seem to capture exactly the right causal relations.

  Evaluate option 1, evaluate option 2, then choose the more valuable

  Note that this would produce a decision even if the mechanisms favouring one or both of the options stopped working just before the decision is taken on the basis of the recorded evaluations.

  Does not have exactly the same causal properties as:

  let option 1 and option 2 simultaneously attempt to influence a decision-maker

  Real, as opposed to simulated neural nets can do this.

Varieties of supervenience

Several varieties of supervenience can be distinguished:

- **property supervenience**: (e.g. having a certain temperature supervenes on having molecules with a certain average kinetic energy);
- **pattern supervenience**: (e.g., supervenience of various horizontal, vertical and diagonal rows of dots on a rectangular array of dots, or the supervenience of a rotating square on changes in the pixel matrix of a computer screen);
- **mereological, or agglomeration, supervenience**: possession of some feature by a whole, arising from a summation of features of parts (e.g. supervenience of the centre of mass of a rock on the masses and locations of its parts, each with its own mass);
- **mathematical supervenience**: e.g. Euclidean geometry can be modelled in arithmetic, using Cartesian coordinates, and in that sense geometry supervenes on arithmetic.
- **mechanism supervenience**: supervenience of one machine on another: a set of interacting objects, states, events and processes supervenes on a lower level reality (e.g., supervenience of a running operating system on the computer hardware).

My topic is mechanism supervenience, relating RVMs to PMs – not the simple case of one property, or entity, relating to others, but a complex *ontology* (collection of diverse entities, events, processes, states, with many properties, relationships and causal interactions) relating to another ontology.
Donald Davidson ("Mental Events", 1970) described supervenience as a relation between properties or "respects", whereas mechanism supervenience involves multiple relations between causally interacting parts and relations of complex ontology-instances, not just properties.

The problems include:

Implications for philosophy, neuroscience, developmental psychology, developmental biology, genetics, information engineering in general, and AI/Robotics – viewed as both science and engineering.

Examples

- Running VMs can be disconnected from input-output interfaces, permanently or temporarily, making conventional means of scientific observation impossible.

  E.g. some problem-solving subsystem may mostly run without connections to sensor or motors and only occasionally produce some result that is transferred to other subsystems. These could justify the cost of the disconnected subsystem.

- Output mechanisms may be incapable of reporting VM processes (e.g. not enough bandwidth).
- Systems using a VM interface for self-monitoring and self control will have timing delays and other discrepancies between processes at the different levels.
- This can lead to misperceptions, during both normal and abnormal operation:

  e.g. Libet effects, phantom limbs, hallucinations, proof-reading errors, psychotic phenomena, mis-descriptions of motivations, etc.

- Much more may be taken in and processed than a self-reporting VM can detect, unless redirected.
- Self-reports in laboratory experiments may report only what the VM architecture makes accessible to the reporting mechanism. (Here be qualia?)
- We need much deeper analyses of varieties of VMs, and different forms of representation, and their possible uses in different subsystems within minds (of animals or machines).

Major problem:

Are there types of virtual machine
that we have not yet thought about
that are required for explaining/replicating
human/animal competences that are currently
beyond the state of the art in AI?

To answer that we have to understand the requirements.

Understanding the requirements requires us to look very closely and analytically at many of the competences shown by humans (of all ages) and other animals both when interacting with the physical environment and when doing other things (e.g. proving theorems in their heads).

For conjectures and evidence about "toddler theorems", see
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#toddler`

For conjectures about types of competition and cooperation in biological VMs not yet available in computer-based VMs see
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk86`

Human and animal vision requires virtual machines

Much research on natural and artificial vision mistakenly assumes that vision is mainly concerned with recognition of objects.

That is a deep error: recognition of objects as normally construed is not required for seeing those objects.

You can see, kick, stroke, push, walk round, climb over and experiment with objects you do not recognise. Compare the Reutersvard cubes shown above.

More importantly, human vision is often perception of processes, not just objects.

Consider what happens when you turn a corner in a busy part of a large town.

There are vast and rapid changes in what is perceived – this cannot involve comparable rapid physical reorganisation of neuronal structures: e.g. rearranging neurons to form a representation of a bus.

But it could involve rapid reorganisation of the contents of an information-processing virtual machine (e.g. patterns of activation).

For examples see the pictures here

`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/multipic-challenge.pdf`

For a discussion of opposing views on the functions of vision see:

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#gibson`
What's vision for, and how does it work? From Marr (and earlier) to Gibson and Beyond

Natural and Artificial Platform VMs

- Platform VMs designed by human engineers provide a basis for constructing new VMs that are implemented in terms of the facilities provided by the platform VM:
  But most such extensions do not arise spontaneously.

  The operating system on your PC can be left running for many years and over time you and others may design and install different software packages, or attach new hardware devices along with device drivers for them. (Compare "hypervisors".)

  If left to itself, a normal operating system will just go on forever waiting for instructions, without initiating any major extensions, though some of them are designed to detect the availability of new versions of old subsystems and download and install them.

- Biological platform VMs, however, are not extended by external designers:
  They have to build and extend themselves

  (partly on the basis of external influences from both the physical environment and conspecifics).

  The requirements to support this have never, as far as I know, been identified.

  The problem is not addressed by research in developmental psychology on which concepts, knowledge or competences are innate.

  Some of the requirements for a "well-designed child" were discussed by McCarthy in this paper `http://www-formal.stanford.edu/jmc/child.html` (written 1996 and published in 2008).

- In humans, biological platform VMs seem to grow throughout infancy and childhood, and for some people (e.g. academics) continue being extended until late in life.

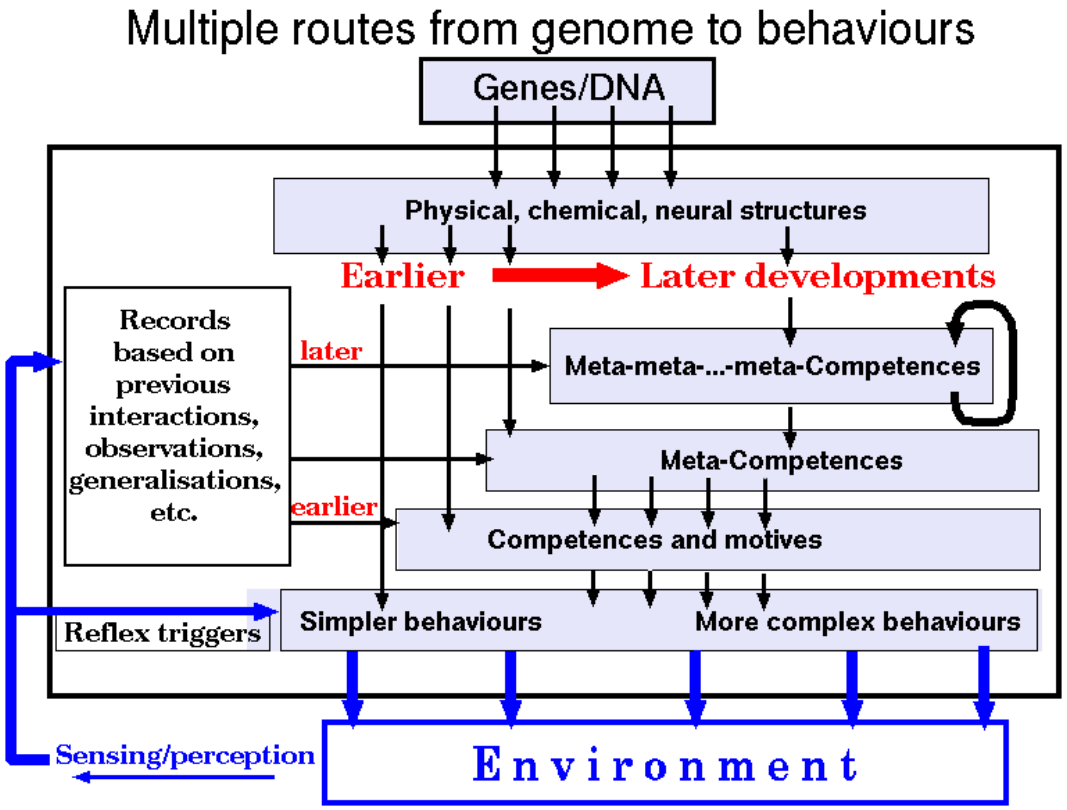
  The extensions support new competences of many kinds, including manipulative and perceptual competences, linguistic, musical and artistic competences, mathematical competences, extended ontologies, new planning and reasoning capabilities, new forms of motivation, new control regimes,

Biological VMs have to grow themselves

Humans and some other species require layered construction of competences and meta-competences in a layered architecture.

Not core competences as normally construed (e.g. by Spelke): rather core architecture-building competences, metacompetences, ...

More like (Karmiloff-Smith, 1992).



The routes from genome to behaviour start earliest on the left, with most direct effects of genome.

The routes to the right are developed later, on the basis of meta-...competences developed after use of earlier competences.

This is work done with Jackie Chappell (IJUC, 2007) – Chris Miall helped with diagram (updated 2015).
"Natural and artificial meta-configured altricial information-processing systems"
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609`,

So far these ideas are conjectures, and have not been implemented in a working system.

RVMs with temporarily or partly 'decoupled' components

A challenge for philosophy of science and psychological methodology.

- "Decoupled" subsystems may exist and process information, even though they have no connection with sensors or motors most of the time.
- Theories referring to them cannot be decisively proved or refuted.
  It takes time to decide which research programme is better:
  Compare Lakatos on methodology of scientific research programmes(Lakatos, 1980)
- For instance, a machine playing games of chess with itself, or investigating mathematical theorems, e.g. in number theory.
- Some complex systems "express" some of what is going on in their VM states and processes through externally visible behaviours.

  However, it is also possible for internal VM processes to have a richness that cannot be expressed externally using the available bandwidth for effectors.

  Likewise sensor data may merely introduce minor perturbations in what is a rich and complex ongoing internal process.

This transforms the requirements for rational discussion of some old philosophical problems about the relationship between mind and body:

E.g. some mental processes need have no behavioural manifestations, though they might, in principle, be detected using 'decompiling' techniques with non-invasive internal physical monitoring.
(This may be impossible in practice, or at best only a matter of partly testable conjecture.
Compare theoretical physics.)

Explaining what's going on in VMs requires a new 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 mostly ignore the variety of complex mappings between RVMs and PMs when discussing causation and when discussing supervenience,

Even though most of them now use multi-process VMs daily for their work.

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 a summary explanation of two kinds of causation (Humean and Kantian) and the relevance of both kinds to understanding cognition in humans and other animals see:

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

Biological Virtual Machines

Biological VMs may be required in plants and animals

- at cellular levels (e.g. control of metabolism)
- in various physiological control mechanisms (e.g. control of balance, running)
- in information-processing architectures of varying complexity
  (e.g. various kinds of perception, learning, motivation, decision-making, ...)

Some biological VMs are pre-specified in the genome, ("precocial VMs") while others ("altricial VMs") are constructed during individual development – in some cases partly under the control of the environment (epigenesis of virtual machines).

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

NOTE

The role of the environment in "controlling" both evolution and individual development implies that the nature of the environment needs to be much more an object of study in developmental psychology and AI than it normally is. (Cf. Ulric Neisser, 1976)

The inside-outside distinction blurs

Often the boundary between machine and environment is different for different sub-systems of the machine.

The physical implementations of some VMs can cross superficial PM boundaries – e.g. VMs that refer to remote, or past, or future entities or events may use external intermediaries to help "tether" (not "ground") the semantic content. (Strawson, 1959)

Different parts of the machine, e.g. different sensors and effectors, may interact with different parts of the environment concurrently.

The machine may treat parts of itself as parts of the environment (during self-monitoring), and parts of the environment as parts of itself (e.g. tools, external memory aids).

See Sloman 1978, chapter 6
`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/chap6.html`

Every mathematician knows that when she is reasoning with the help of diagrams, logical formulae or equations, on paper or on a blackboard part of her mind is physically external to her body.

Some philosophers over-impressed by embodiment think this is a new idea. It's old.

I remember as a student in the 1960s hearing discussions of a blind person's stick, a dentist's probe, and diagrams used for reasoning. See also P.F. Strawson's (1959) referential causal chains and J.J. Gibson's (1966) work on active perception.

There's a general problem of lack of knowledge of what has already been said and written, getting steadily worse: The "Singularity of Cognitve Catchup" (SOCC) `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/another-singularity.html`

We still don't know how to put ghosts with animal intelligence into machines

But we can make some suggestions about where to look for answers.

Every intelligent ghost must contain
An **information processing** machine

It's a very long term research programme.

Matter, Energy and Information

A related talk was given at SAB2010 at a museum of Leonardo da vinci's work at Clos-Lucé, Amboise.

On the bus from Paris we had a video presentation on Leonardo's life and work.

One of things I learnt was that he had written: "Everything is motion."

He seems to have been saying that the world can be seen as matter and energy interacting and constantly changing features and relationships.

I suspect that if he were alive today, he would agree that

Besides matter and energy there is information – all embedded in space and time.

and

Besides many varieties of motion there are many varieties of information-processing

Our understanding of that variety is still in its early stages – only since the 20th century have we begun to understand how to make anything but the very simplest information processing machines.

There were some precursors, e.g. Charles Babbage and Ada Lovelace – who anticipated some of Turing's ideas about the significance of computers.

An important feature of information processing is that it can be a form of control: information can specify what to do, how to do it, how to weigh up conflicting alternatives, and many more – These are major aspects of animal information processing.

A framework for developing these ideas

John Maynard Smith and Eörs Szathmáry (1995) proposed that there are eight major transitions in evolution, summarised here:

`http://en.wikipedia.org/wiki/The_Major_Transitions_in_Evolution`

Transition from:	Transition to:
1 Replicating molecules	"Populations" of molecules in compartments
2 Independent replicators (probably RNA)	Chromosomes
3 RNA as both genes and enzymes	DNA as genes; proteins as enzymes
4 Prokaryotes	Eukaryotes
5 Asexual clones	Sexual populations
6 Protists	Multicellular organisms - animals, plants, fungi
7 Solitary individuals	Colonies with non-reproductive castes
8 Primate societies	Human societies with language, enabling memes

They identified features common to the eight transitions:

1. Smaller entities come together to form larger entities.
2. Smaller entities become differentiated as part of a larger entity.
3. Smaller entities become unable to replicate in the absence of the larger entity.
4. Smaller entities become able to disrupt the development of a larger entity.
5. **New ways arise of transmitting information.**

   Is that the only kind of information-processing transition?

Suggestion: We should try to produce a taxonomy of types of evolutionary or developmental transition (gradual or discontinuous) connected with information processing – in animals, and in future animats.

Besides information transmission there is also manipulation.

(A century or two should suffice.)

Another set of transitions

We can generalise the last feature to include

- new kinds of information (new information contents)
- new forms of representation (new physical and virtual media)

  For a discussion of some of the language-like forms of representation that must have evolved in other animals for internal use, and must be present in human children before they begin to speak

  (e.g. it is required for visual and other forms of perception, for wanting, for trying, for planning and executing actions, and in order to have a need or a desire to communicate)

  see `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang`
  Evolution of minds and languages. What evolved first and develops first in children: Languages for communicating, or languages for thinking (Generalised Languages: GLs)
- new ways of acquiring, processing, or using information.
- new information-processing architectures

These changes can be related to different sorts of trajectories in evolution and in development.

Trajectories in different spaces

- Trajectories in the space of possible sets of requirements (niches)
- Trajectories in the space of designs for systems satisfying requirements
- Trajectories in The space of implementations for each design

Transitions in information processing

Work to be done:

We need to show how some of the transitions in what we have designed and built in complex computing systems can suggest a strategy for helping Darwin answer his critics who think natural selection could not have produced mental processes.

A full answer requires studying

- Evolutionary transitions in requirements for information processing in organisms (their niches)
- Evolutionary transitions in designs and implementations
  (Including development of new mechanisms for use in implementations.)

  Especially evolution of the use of running virtual machines of various kinds

  As pointed out in the conference: evolutionary computation may one day be used to simulate some of these evolutionary processes.

  But it will have to be a much richer form of evolutionary computation than anything tried so far: with interacting evolutionary and developmental trajectories in many species interacting on a fairly large planet.

More on what needs to be explained.

There is much more to be said about requirements for virtual machines that could be adequate to explain everything that is already known about human and animal minds.

Different researchers focus on different subsets of phenomena,

- partly because they study different species
- partly because they are interested in explaining different things
- partly because they are interested in different applications
- partly because they accept different positions in philosophy of science

In order to make deep progress we need to be very broad-minded about what needs to be explained so that when we build models that explain a restricted subset we are well aware of what has NOT been achieved thereby.

For example if we wish to explain the full range of human phenomena about consciousness then we need to explain how individual philosophers trying to explain consciousness can get confused and arrive at different conclusions: we need a design that can accommodate such diverse types of philosophical thinking and puzzlement.

Example: Zen and the Art of Consciousness

Some unusual examples related to meditation are discussed in Susan Blackmore's little book **Zen and the Art of Consciousness** (Blackmore, 2011)

A partial review is here:
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/blackmore-zen-consciousness.html`

- One of the recurring themes in that book is the difficulty of specifying precisely what the contents of consciousness are, even when one is trying to do nothing but attend to the contents.
- It is also hard to specify exactly at what time an item enters consciousness, or what was going on before an enduring feature is attended to,

  e.g. a background noise or a pattern on the carpet, or leaves moving in the wind.
- I think the key idea is that there is no binary division between being in or out of consciousness, not because it is a matter of degree, but because:
  - what is in your consciousness is information that is available to you for various potential uses (reporting, remembering, painting, using to control action, evaluating, etc.)
  - availability is usually **conditional** either because what is available depends on which active processes occur (e.g. switching attention) or because it depends other factors (e.g. the volume of a sound, or what other sounds exist at the same time, or the speed of motion, or the contrast with the visible background.) that may themselves be conditional on other things.
  - some things are more directly/immediately available because most of the conditions are already satisfied, whereas others are more or less remote because additional conditions need to be satisfied before the information can be accessed.
- More research, and more experimentation with working models is needed.

References

These slides present a small part of a large picture, elaborated in papers and presentations listed below – though there is still much to be done (e.g. (Sloman, 2010d)).

References

Blackmore, S. (2011). *Zen and the Art of Consciousness*. Oneworld Publications. (Previous title: Ten Zen Questions)

Chappell, J., & Sloman, A. (2007). Natural and artificial meta-configured altricial information-processing systems. *International Journal of Unconventional Computing*, *3*(3), 211–239. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/07.html#717

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Lakatos, I. (1980). Falsification and the methodology of scientific research programmes. In J. Worrall & G. Currie (Eds.), *Philosophical papers, Vol I* (pp. 8–101). Cambridge: Cambridge University Press.

Maynard Smith, J., & Szathmáry, E. (1995). *The Major Transitions in Evolution*. Oxford, England:: Oxford University Press.

Sloman, A. (1971). Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd ijcai* (pp. 209–226). London: William Kaufmann. Available from http://www.cs.bham.ac.uk/research/cogaff/62-80.html#1971-02 (Reprinted in *Artificial Intelligence*, vol 2, 3-4, pp 209-225, 1971)

Sloman, A. (1978). *The computer revolution in philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press). (http://www.cs.bham.ac.uk/research/cogaff/62-80.html#crp, Revised 2015)

Sloman, A. (1993). The mind as a control system. In C. Hookway & D. Peterson (Eds.), *Philosophy and the cognitive sciences* (pp. 69–110). Cambridge, UK: Cambridge University Press. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#18

Sloman, A. (2009a). Architecture-Based Motivation vs Reward-Based Motivation. *Newsletter on Philosophy and Computers*, *09*(1), 10–13. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/misc/architecture-based-motivation.html

Sloman, A. (2009b). Machines in the Ghost. In D. Dietrich, G. Fodor, G. Zucker, & D. Bruckner (Eds.), *Simulating the Mind: A Technical Neuropsychoanalytical Approach* (pp. 124–148). Vienna & New York: Springer. Available from http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0702

Sloman, A. (2009c). Some Requirements for Human-like Robots: Why the recent over-emphasis on embodiment has held up progress. In B. Sendhoff, E. Koerner, O. Sporns, H. Ritter, & K. Doya (Eds.), *Creating Brain-like Intelligence* (pp. 248–277). Berlin: Springer-Verlag. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#912

Sloman, A. (2010a). An Alternative to Working on Machine Consciousness. *Int. J. Of Machine Consciousness*, *2*(1), 1–18. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#910

Sloman, A. (2010b, August). How Virtual Machinery Can Bridge the "Explanatory Gap", In Natural and Artificial Systems. In S. Doncieux & et al. (Eds.), *Proceedings SAB 2010, LNAI 6226* (pp. 13–24). Heidelberg: Springer. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/10.html#sab

Sloman, A. (2010c). Phenomenal and Access Consciousness and the "Hard" Problem: A View from the Designer Stance. *Int. J. Of Machine Consciousness*, *2*(1), 117–169. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#906

Sloman, A. (2010d). *Supervenience and Causation in Virtual Machinery*. University of Birmingham, School of Computer Science. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk86 (Online tutorial presentation)

Sloman, A. (2010e). *Talk 85: Daniel Dennett on Virtual Machines.* University of Birmingham, School of Computer Science. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk85 (Online tutorial presentation)

Sloman, A. (2010f). *Using virtual machinery to bridge the "explanatory gap"; Or: Helping Darwin: How to Think About Evolution of Consciousness; Or: How could evolution (or anything else) get ghosts into machines?* University of Birmingham, School of Computer Science. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk84 (Invited keynote talk at SAB2010)

Sloman, A. (2011a). Varieties of Meta-cognition in Natural and Artificial Systems. In M. T. Cox & A. Raja (Eds.), *Metareasoning: Thinking about thinking* (pp. 307–323). Cambridge, MA: MIT Press. Available from http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0803

Sloman, A. (2011b). What's information, for an organism or intelligent machine? How can a machine or organism mean? In G. Dodig-Crnkovic & M. Burgin (Eds.), *Information and Computation* (pp. 393–438). New Jersey: World Scientific. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#905

Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, *10*(4-5), 113–172. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200302

Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. London: Methuen.