Why the "hard" problem of consciousness is easy and the "easy" problem hard.

(And how to make progress)

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

These PDF slides will be available in my 'talks' directory:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#cons09

Why? Because:

1. The “hard” problem can be shown to be a non-problem because it is formulated using a seriously defective concept (explained later as the concept of “phenomenal consciousness” defined so as to rule out cognitive functionality).
2. So the hard problem is an example of a well known type of philosophical problem that needs to be dissolved (fairly easily) rather than solved.

   For other examples, and a brief introduction to conceptual analysis, see
   `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/varieties-of-atheism.html`
3. In contrast, the so-called “easy” problem requires detailed analysis of very complex and subtle features of perceptual processes, introspective processes and other mental processes, sometimes labelled “access consciousness”: these have cognitive functions, but their complexity (especially the way details change as the environment changes or the perceiver moves) is considerable and very hard to characterise.
4. “Access consciousness” is complex also because it takes many different forms:

   what individuals can be conscious of, and what functions being conscious has, varies hugely, from simple life forms to sophisticated adult humans, and can vary between humans at different stages of development from conception to senile dementia. The concept is highly polymorphic.
5. Finding ways of modelling these aspects of consciousness, and explaining how they arise out of physical mechanisms, requires major advances in the science of information processing systems – including computer science and neuroscience.

The remaining slides attempt to justify these claims

Some of the topics included, not necessarily in this order

1. Philosophical (and psychological) background to the idea of 'phenomenal consciousness' and qualia as the introspectable contents of consciousness.

2. Brief introduction to philosophers' notions of supervenience/realisation/implementation.

3. Ned Block's distinction between access consciousness (which has a functional role) and phenomenal consciousness (which does not).

4. Why there seems to be an 'explanatory gap' (T.H.Huxley) or a 'hard problem' (Chalmers). Chalmers uses something close to Block's notion of phenomenal consciousness to pose the problem, but it is actually a very old problem.

   [Incidentally the explanatory gap was used even by scientists sympathetic to Darwin, as an objection his theory of evolution as applied to humans.]

5. Chalmers uses the gap to propose a research programme investigating 'Neural correlates of consciousness'. I'll try to show why the notion of phenomenal consciousness (not access consciousness) is incoherent, as is the search for neural correlates.

6. Brief introduction to the technology of virtual machinery (not virtual reality) developed over the last six decades and its philosophical relevance: processes and events in running virtual machines can be causes and effects, despite being fully implemented in physical mechanisms. (But specific VM entities, events, and processes, need not have physical correlates.)

7. Explanation of the importance of virtual machines in sophisticated control systems with self-monitoring and self-modulating capabilities. Why such machines need something like "access consciousness"/qualia – and why they too generate an explanatory gap – a gap bridged by a lot of sophisticated hardware and software engineering developed over a long time.

   Here the explanations are much deeper than mere correlations: we know **how** the physical and virtual machinery are related.

8. Conjecture that biological evolution discovered those design problems long before we did and produced solutions using virtual machinery long before we did – in order to enable organisms to deal with rapidly changing and complex information structures (e.g. in visual perception, decision making, control of actions, etc.).

You can't rapidly rewire millions of neurons when you look in a new direction. So there's no alternative to using virtual machinery?

9. Some deep philosophical problems about causation, and about meaning/intentionality, discussed briefly, using simple (perhaps over-simple) examples.

9. Work still to be done, e.g. finding out precisely what the problems were that evolution solved and how they are solved in organisms, and how things can go wrong in various ways, and why future intelligent robots will need similar solutions.

Preamble (Historical caricature)

- The problem of explaining how mind and matter are related goes back at least two millenia.
  `http://plato.stanford.edu/entries/aristotle-psychology/`
- The usefulness of talking about consciousness has been much debated by scientists,
  During the 20^{th} C. many (English speaking) psychologists tried to avoid talking about anything mental.
- Freud's theories with his division between id, ego, and superego were disparaged
  (though welcomed by many outside psychology departments).
- The rise of cognitive psychology and cognitive science since late 1960s, along with development of computational theories began to change this.
- More hard-headed scientific psychologists started talking about emotions in the 1980s.

  The journal *Cognition and Emotion* started in 1987. Now there are journals, workshops, projects, ...
- In the 1990s people started talking about "neural correlates of consciousness"(NCCs)

  "Progress in addressing the mind-body problem has come from focusing on empirically accessible questions rather than on eristic philosophical arguments. Key is the search for the neuronal correlates - and ultimately the causes - of consciousness." (Scholarpedia) (Compare Crick & Koch 1990)

  see `http://www.scholarpedia.org/article/Neural_correlates_of_consciousness`
- Neural correlates of what??? Some respond: "Phenomenal consciousness".

  In 1995 Ned Block distinguished access consciousness (A-C) (introspective states with cognitive functions) and phenomenal consciousness (P-C) (states without cognitive functions).
- David Chalmers *The Conscious Mind: In Search of a Fundamental Theory* (1996) claimed that there was a "hard" problem of consciousness, namely explaining the existence of P-C in a physical world.

  Explaining A-C was thought to be (relatively) easy by comparison.
- This encouraged many to start searching for NCCs as a way of attacking the "hard" problem. What is the hard problem?
- Now many AI/Robotics papers are sprinkled with references to "emotion" and "consciousness" – mostly undisciplined and muddled as far as I can see.

Next few slides

- There are many empirical facts about human experience (most of them easily checked) that support claims about the existence of introspectively accessible entities, often described as privately accessible contents of consciousness.

  Various labels are used for these entities: “phenomenal consciousness”, “qualia” (singular “quale”), “sense-data”, “sensibilia”, “what it is like to be/feel X” (and others).
  For a useful, but partial, overview see `http://en.wikipedia.org/wiki/Qualia`

- What is not clear is what exactly follows from the empirical facts, and how best they can be described and explained.

- I start by indicating some of the standard philosophical problems and theories about the status of the entities allegedly shown to exist.

- They pose a scientific problem of explaining how such entities arise and how they are related to non-mental mechanisms, e.g. brains.

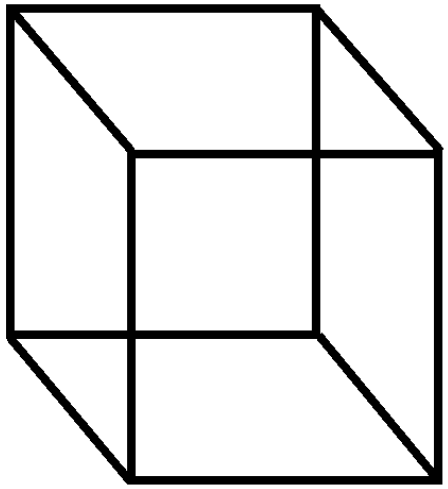
- I introduce and then criticise the so-called “hard problem”, which is a new formulation of what was long ago called “the explanatory gap”

- Later I present a new way of dealing with the explanatory gap using the concept of “running (or active) virtual machine”.

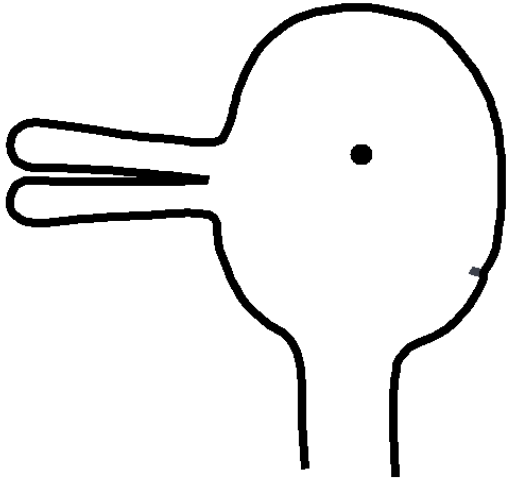
  This concept has been largely ignored (or when not ignored, misunderstood) by philosophers and psychologists even though they use examples every day when writing papers, reading and writing research papers, using spelling checkers, spreadsheets, internet browsers, anti-virus packages, etc.

Do some "consciousness science"

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 paper 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".

If the figures do not flip for you, ask a friend to try.

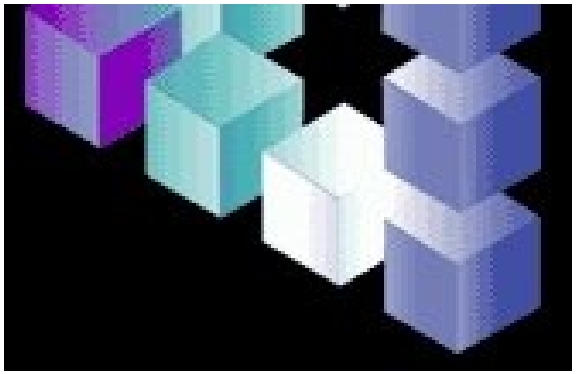
Empirical basis for referring to contents of consciousness

Introspection provides empirical evidence for mental contents "inside" us, distinct from external causes or internal physical processes:

- Ambiguous figures: as you stare at them, nothing changes out there, only the experiences/qualia (in your mind) "flip" (E.g. Necker cube, face-vase, old/young lady, etc.)
  This one rotates in 3-D in either direction: `http://www.procreo.jp/labo/labo13.html`
- Optical illusions (of many kinds): Muller-Lyer, Ebbinghaus, motion/colour after-effects.
- Dreams, hallucinations, hypnotic suggestions, effects of alcohol and other drugs.
- Different people see the same things differently. E.g. short-sighted and long-sighted people.
  Identical twins look different to people who know them well, but look the same to others.
  Cf. Botanical expertise makes similar plants look different. Colour blindness.
- Pressing eyeball makes things appear to move when they don't, and can undo binocular fusion: you get two percepts of the same object; crossing your eyes can also do that.
- Put one hand into a pail of hot water, the other into a pail of cold water, then put both into lukewarm water: it will feel cool to one hand and warm to the other. (A very old philosophical experiment.)
- People and other things look tiny from a great height – without any real change in size.
- Aspect ratios, what's visible, specularities, optical flow – all change with viewpoint.
- We experience only portions of things. A cube has six faces but we can't see them all: what you experience changes as you move.
- Thinking, planning, reminiscing, daydreaming, imagining, can be done with eyes closed ....
- Composing poems, or music, or proving theorems with your eyes shut.
  Rich mental contents (not perceptual contents) can be involved in all of these.

2-D and 3-D Qualia (added 27 Jan 2010)

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

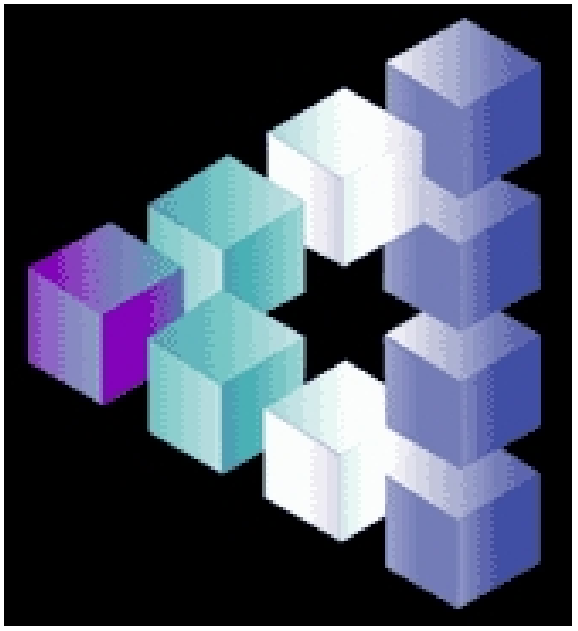


As with the Necker cube you have experiences of both 2-D lines, regions, colours, relationships and also 3-D surfaces, edges, corners, and spatial relationships.

You probably also experience various affordances: places you could touch the surfaces, ways you could grasp and move the various cubes (perhaps two 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 on the opposite side.

The second picture on the right (from which the first 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.

But the scene depicted in the lower 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 be inconsistent!

Implications of the empirical facts

When we perceive our environment or when we think, plan, daydream, reminisce, compose in our heads, in addition to what's "out there", or what's happening in our brains, there are mental things going on in us, which we can attend to, think about, ask questions about, and in some cases use, e.g. to gain information about the environment.

- The process of discovery of these internal entities mainly uses introspection.
  Introspection is a kind of perception, directed inwardly, but is inherently "private".
- Contents of introspection (what we experience as existing in us) are what discussions of phenomenal consciousness and the hard problem of consciousness are supposed to be about. Terminology varies: Qualia, sense-data, ideas, impressions, sensibilia, ...
- A science of mind needs to study of these phenomena, even though they are not accessible by the normal methods of science: "public" observation and measurement.
- The internal states, events and processes (desires, decisions, impulses, "acts of will"...) are also able to produce physical events and processes
  e.g. movements, speech, saccades, smiling, weeping, ...
- The internal entities are distinct from physical matter in many ways:
  e.g. they are not amenable to observation and measurement by physicists, chemists, etc. using the tools of the physical sciences – they are only accessible by their owner.
  Even if you could somehow be connected to my brain so as to have my experiences projected into you, you would then be having your experiences through my senses – you would not be sharing my experiences. You still could not know for certain, as I can, what my experiences actually are.
- Can we bridge the "explanatory gap": between physical phenomena and contents of consciousness, in either direction?

Questions generated

Difference between mental contents and physical phenomena generate many questions. Some examples:

- Epistemological:
  - Can we infer the existence of an "external" world from our experiences?
  - Can we infer the existence of "other minds" from our experiences
  - Can we tell whether infants, other species, or robots, have them? ... etc....
- Metaphysical:
  - What kinds of things exist and how are they related?

    Can mind and matter influence each other? Can either exist without the other?
- Conceptual/semantic:
  - How can we conceive of and think about things we don't directly experience?

    (e.g. the physical environment, other minds, causal connections, distant past or future events....)
  - Is there some other way of acquiring concepts other than by abstraction from experiences of instances?

    Concept empiricism (recently reinvented as "symbol grounding theory") says no. But Kant and 20th century philosophers of science showed that concept empiricism must be false: there must be other ways of acquiring concepts, e.g. of gene, subatomic particle, remote galaxy, etc.
- Scientific:
  - Can we explain how physical systems can produce mental contents?

    E.g. what brain mechanisms could do it?
  - Can we account for the evolution of minds with experiences in a physical world?
  - Can we understand development of individual minds (e.g. from foetus to philosopher?)

Answers suggested by various thinkers

A (partly) new collection of answers (some going back a very long time):

- Epistemological: (We need to remember our biological origins.)
  - We don't infer an "external" world: we evolved so as to assume that we exist in a 3-D space and able to find out what sorts of things are in it and how they work.
  - Likewise evolution, not inference, makes us assume there are "other minds".

  What work is done by those two assumptions, and how, is another matter: discussed later.
- Metaphysical:
  - Many kinds of things exist, with complex interactions: science keeps finding more.
- Conceptual/semantic:
  - As Kant argued, you can't have experiences without concepts, so concepts don't all come from experience.

    Scientific concepts have to be introduced through the theories that use them. We are still learning (very slowly) how to give machines such capabilities.
    `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#grounding`
  - Recently developed sets of concepts used in designing and implementing virtual machines whose behaviours are very different from physical processes, give us new conceptual tools for thinking about minds. (More on this later.)
- Scientific:
  - By learning how to replicate the creation of biological minds as virtual machines running on, and controlling, physical machines, we may become more ready to answer the really hard scientific questions and close the explanatory gap.

Plato's cave

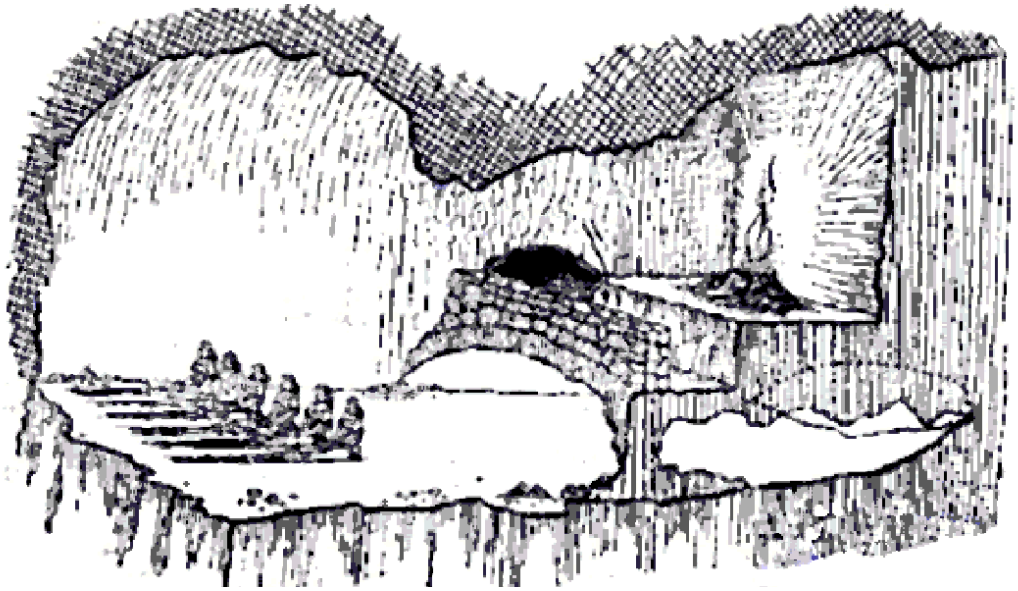
Plato's cave analogy expresses the view that we are like prisoners in a cave who do not see things in the world as they are, since all we have are the shadows cast on a wall of the cave (the contents of consciousness).

http://home.earthlink.net/~johnrpenner/Articles/PlatosCave.html

If people walk past the flame on the raised pathway, their shadows will be cast on the wall, and be seen by the prisoners facing the wall.

The prisoners cannot turn their heads to see the things that cast the shadows.

This could be a metaphorical description of the relationship between our percepts and their causes in the environment.
(But I am no Plato Scholar.)



Picture found in many locations on the internet

But if the prisoners are scientists, they may be able to develop, and perhaps also test, theories about the causes of the shadows.

E.g. explaining what happens to some shadows when they move their own heads?

Older theories about mind-body relations

Various attempts have been made to describe the relations between:

(a) physical objects, including bodies, sense organs and brains, and

(b) mental entities, e.g. minds and their contents.

An incomplete sample:

- Monism: There is only one kind of stuff that exists
  Monists differ as to whether the stuff is physical, mental, or something else:
  - Idealism (everything is mental): to be is to be experienced (Berkeley: "Esse est percipi")
  - Phenomenalism: existence is enduring possibility of being experienced (e.g. Hume?)
  - Materialism: only matter exists, and some portions of matter behave as if perceiving, desiring, thinking, etc.
  - Behaviourism: mental phenomena just **are** behaviours or behavioural dispositions, tendencies, capabilities.
  - Double-aspect theories
    * There is only one kind of stuff, but it has different aspects (neutral monism)
    * Mind-brain identity – mental processes just are the physical processes that implement them.
- Dualism:
  - Interactionism: Two kinds of stuff exist, and interact causally
  - Psycho-physical parallelism: Two streams existing independently – with no causal connections
  - One way causation – e.g. matter to mind (Epiphenomenalism), or mind to matter (???-ism).
- Polyism:(??) There are **many** kinds of stuff interacting in different ways.
- Expressivism: labelling something as conscious expresses a stance or attitude to it.
  E.g. Daniel Dennett's "intentional stance". A stance is useful, or useless, not true or false.

A-C and P-C

All of the above leave many people feeling that there is an "explanatory gap" between the physical and the mental.

Ned Block's 1995 paper made an influential distinction between access consciousness (A-C, which has a functional role) and phenomenal consciousness (P-C, which does not).

`http://cogprints.org/231/0/199712004.html`

"The information processing function of phenomenal consciousness in Schacter's model is the ground of what I will call "access-consciousness". A perceptual state is access-conscious roughly speaking if its content – what is represented by the perceptual state – is processed via that information processing function, that is, if its content gets to the Executive system, whereby it can be used to control reasoning and behavior."

"Let me acknowledge at the outset that I cannot define P-consciousness in any remotely non-circular way. I don't consider this an embarrassment."

"The controversial part is that I take P-conscious properties to be distinct from any cognitive, intentional, or functional property. (Cognitive = essentially involving thought; intentional properties = properties in virtue of which a representation or state is about something; functional properties = e.g. properties definable in terms of a computer program"....) (Note: he should not have referred to computer programs: there is a deeper, broader notion of function involving causal role – See quotes from Chalmers, below.)

"It is of course P-consciousness rather than access-consciousness or self-consciousness that has seemed such a scientific mystery."

Notice that P-C is defined negatively: e.g. it has no functional role, unlike A-C.

This implies that P-C has no cognitive causal powers.

This is a bit like postulating invisible, intangible, fairies at the bottom of the garden produced by the plants growing, but incapable of influencing anything else.

How can things with no cognitive function generate so much philosophical questioning???

More on gaps

Chalmers (see next slide) used a distinction very close to Block's distinction between A-C and P-C, to pose "the hard problem of consciousness" which actually is the very old problem, expressed thus by T.H. Huxley (1866):

> "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 Djin when Aladdin rubbed his lamp."

Many scientists around Huxley's time thought Darwin's arguments for evolution of animal bodies were convincing in the light of evidence of small changes in physical structure, but also felt that evolution could not produce anything like states of consciousness as a result of a succession of gradual changes in physical matter. (Huxley's "explanatory gap".)

Chalmers suggests there are two gaps, which we can describe in Block's terminology:

- a gap separating physical mechanisms from access consciousness
- a gap separating physical mechanisms from phenomenal consciousness

He thinks bridging the first gap ("the easy problem") is non-trivial, but much easier than bridging the second: "the hard problem of consciousness".

He proposes a research programme investigating 'Neural correlates of consciousness' in order to understand the second gap better.

But if phenomenal consciousness is defined as incapable of having any cognitive functions, then I shall argue that the concept is incoherent, in which case we can solve the hard problem easily by showing that it is a non-problem: just a philosophical muddle.

Explaining how A-C can exist in a physical world remains hard, but tractable, if we use the right tools.

Explanatory gaps and the hard problem

David Chalmers re-labelled the 'explanatory gap' (T.H.Huxley, 1866) as 'the hard problem of consciousness', which can be seen as the problem of explaining how physical machinery can produce P-C.

He does not mention Block or Huxley, but essentially makes the same distinction as Block between A-C and P-C, in different words:

`http://www.imprint.co.uk/chalmers.html`
Facing Up to the Problem of Consciousness (JCS, 1995)

"The easy problems of consciousness include those of explaining the following phenomena:

the ability to discriminate, categorize, and react to environmental stimuli;
the integration of information by a cognitive system;
the reportability of mental states;
the ability of a system to access its own internal states;
the focus of attention;
the deliberate control of behavior;
the difference between wakefulness and sleep.

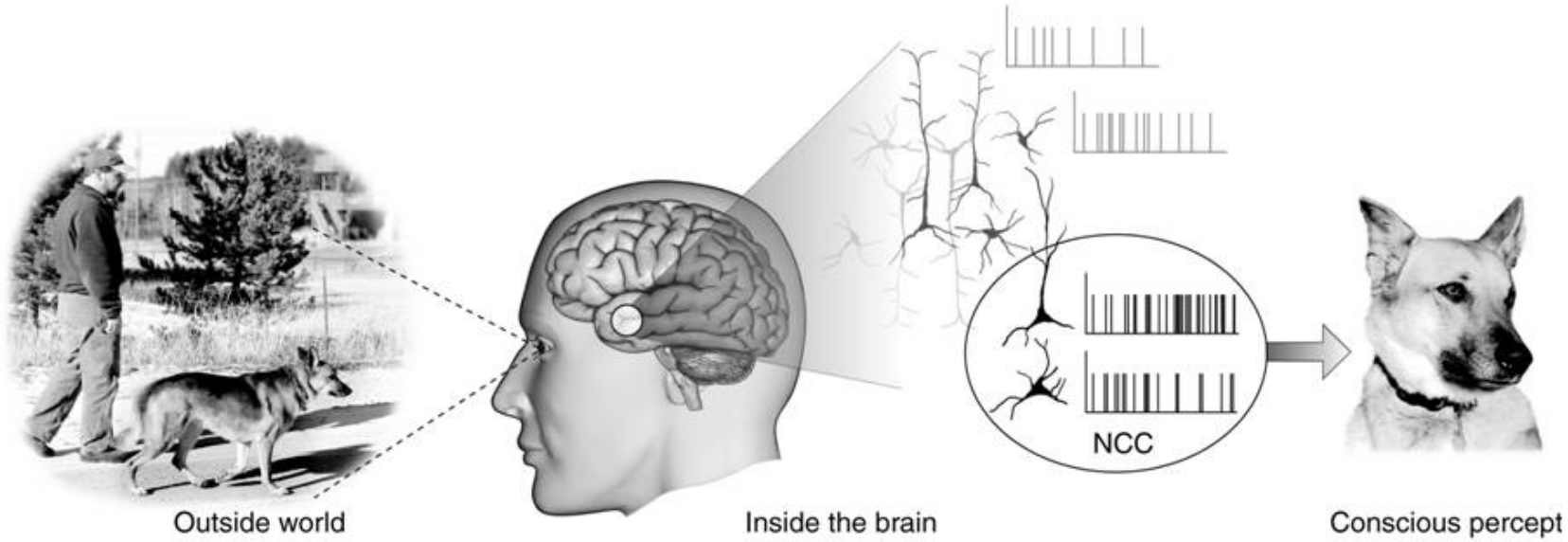
.... "There is no real issue about whether these phenomena can be explained scientifically. All of them are straightforwardly vulnerable to explanation in terms of computational or neural mechanisms. .... Getting the details right will probably take a century or two of difficult empirical work." ...
"If any problem qualifies as the problem of consciousness, it is this one. In this central sense of 'consciousness', an organism is conscious if there is something it is like to be that organism, and a mental state is conscious if there is something it is like to be in that state. Sometimes terms such as 'phenomenal consciousness' and 'qualia' are also used here, but I find it more natural to speak of 'conscious experience' or simply 'experience'...."

He is making essentially the same distinction as Block (using different labels) and proposing that explaining how physical systems can produce P-C is the "hard problem".

Figure for Neural Correlates of Consciousness

by Mormann & Koch

From Scholarpedia



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

Problem: does causation go only one way?

How can motives, decisions, intentions, plans, preferences.... produce behaviour?

Or is it a myth that they do?

By definition, P-C cannot do that. What about A-C? How?

What it is like to be X

I find it astounding that so many thinkers treat the phrase “what it is like to be” as having sufficient precision to play a role in defining a serious problem for philosophy or science.

It was originally brought into philosophical discussion by Tom Nagel in “What is it like to be a bat?” in 1970

Compare my “What is it like to be a rock?”

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

As a piece of colloquial language the phrase is used in different contexts to express different things. E.g. there are great differences between what it is like to be

- half asleep
- horizontal
- unpopular
- 25 years old
- in France
- able to cook a meal without help
- drowning
- deaf
- more knowledgeable than most of the population
- unwittingly on the verge of a great discovery
- dead
- unknown

The phrase expresses a polymorphic concept: what is being said depends on the whole context in which it is being used, and there need not be anything in common to all the possible interpretations of “what it is like to be” – as with the word “conscious”.

There is another way: notice that minds **DO** things

What many people forget when they discuss these issues is that minds don't just contain bits of consciousness: they contain **processes** and those processes **do things**, including producing, modifying and aborting, other processes, mental and physical.

- In other words, minds are machines and the philosophical and scientific task is to find out what they do, and what sorts of physical and non-physical machinery they use.

  This is deeper and harder than asking what things minds contain, and how those things are produced, and what non-mental things correlate with them.

- Intelligent minds produce and use theories, questions, arguments, motives, plans, percepts: all of which are information structures (including descriptive and control structures).
- Such minds require information-manipulating machinery.

  Some people may think their minds contain only sequences of sentences and pictures, or sentence-like and picture-like entities.
  But sentences and pictures are merely two sorts of formats for encoding various kinds of information.
  A mind that understands sentences and pictures must have more than sentences and pictures, since if understanding sentences and pictures amounted to having and understanding more sentences and pictures, that would require an infinite regress of sentences and pictures.

- The implications of all that tend to be ignored in many discussions of consciousness.
- **Most researchers investigating these topics lack powerful conceptual tools developed in computer science/software engineering for creating, analysing, understanding information processing machinery of various kinds.**

Information processing models of mind

Some researchers have attempted to explain what consciousness is in terms of information-processing models of mentality, and they often assume that such models could be implemented on computers.

Several decades ago, Ulric Neisser in "The Imitation of Man by Machine" *Science* (1963)
`http://www.sciencemag.org/cgi/reprint/139/3551/193`
distinguished what Robert Abelson later called "cold cognition" (perception, reasoning, problem solving, etc.) and "hot cognition" (desires, emotions, moods, etc.).

Neisser claimed that the former could be replicated in computers but not the latter, offering detailed examples to support his case.

Herbert Simon responded to this challenge in "Motivational and emotional controls of cognition" *Psychological Review* 1967 `http://circas.asu.edu/cogsys/papers/simon.emotion.pdf`
arguing that computation can include *control* as well as "cold" cognition.

The idea was that emotions and other forms of "hot cognition" could be analysed as states and processes produced by control mechanisms, including motive generators and interrupt mechanisms.

My own earliest attempts were in *The Computer Revolution in Philosophy* 1978, e.g. chapters 6–10.
`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`
Similar ideas were put forward before that and after that, e.g. in P.N. Johnson-Laird's book *The Computer and the Mind: An Introduction to Cognitive Science* (1988)

People objecting to such ideas claim that machines with all the proposed forms of information processing could still lack consciousness: no matter how convincing their behaviour, they could be mere "zombies" (defined in the next slide).

What's a zombie?

Many of the discussions of consciousness by philosophers refer to zombies: e.g. claiming that no matter how closely you build a machine that looks and behaves like a human and which contains a physical information-processing machine mimicking all of the functionality of a human brain it could still be a zombie.

The idea of a zombie occurs in a certain kind of grisly science fiction.

Here are some definitions:

- "A zombie is a creature that appears in folklore and popular culture typically as a reanimated corpse or a mindless human being. Stories of zombies originated in the Afro-Caribbean spiritual belief system of Vodou, which told of the people being controlled as laborers by a powerful sorcerer. Zombies became a popular device in modern horror fiction, largely because of the success of George A. Romero's 1968 film Night of the Living Dead."
- "zombi: a dead body that has been brought back to life by a supernatural force"
- "zombi: a god of voodoo cults of African origin worshipped especially in West Indies"
- "automaton: someone who acts or responds in a mechanical or apathetic way"

Many philosophers have come to regard the explanatory gap as a challenge to demonstrate that certain sorts of physical machines could not be zombies: can we find a design that **guarantees** that its instances have minds and are conscious, and do not merely **behave** as if they were conscious.

However if being conscious means having P-C that's an incoherent challenge!

(As explained below.)

Spurious philosophical arguments: imaginability

Dualists and idealists have been convinced that they can think about the mental as distinct from the physical/material in a manner that refutes various explanations of what mental phenomena are, as follows:

- Whatever a theorist proposes as constituting or explaining the existence of mental phenomena (e.g. certain collections of brain processes, or certain collections of internal cognitive processes and dispositions), such a dualist responds:

  "I can **imagine** a machine that has all of that and behaves exactly like a conscious human, but is really a zombie: it has no consciousness – it does not have this" said pointing inwardly.

- Philosophers who can imagine a zombie made according to specifications XYZ often claim that that **proves** that XYZ cannot provide an analysis of what consciousness is, or explain how it arises.
- The A-C/P-C distinction is sometimes appealed to in such arguments: saying that such a design can produce A-C, but not P-C, not the "what it is like to be" states. So the design does not explain P-C.
- Such imaginability/conceivability of machinery without consciousness proves nothing when the machine specification is rich and detailed enough to provide the basis for a design for a working system that goes through all the internal cognitive processes required for functioning animals, including: interpretation of sensory inputs, reasoning, learning, remembering, imagining, generating motives, deliberating, deciding, controlling actions, and monitoring internal processes. (We are nowhere near such a design.)

When enough such detail has been achieved, claiming that there is something that has all that richness but lacks contents of consciousness is a form of self delusion, **for what is claimed to be imagined is actually incoherent, as I'll try to explain. In other words,** the concept of functionless, causally disconnected contents of consciousness (P-C) is incoherent.

Beware of arguments from imaginability

People can **think** they are imagining something while fooling themselves – because what they are referring to is either totally impossible or else semantically detached from reality, or mathematically impossible.

- Someone may claim to be able to imagine it being 3am on the surface of the sun, or noon at the centre of the earth.

  But "the time at X" depends on angular elevation of the sun above the horizon at X.
  Where there is no angular elevation, talk about "The time at X" makes no sense.

- Another old example is the claim to imagine the whole universe moving slowly in an north-west direction at three miles per hour, or being able to imagine everything in the universe expanding at a uniform rate, so that the expansion is not detectable.

  In both those cases the claim is incoherent because what is imagined is defined so as to undermine preconditions for what is imagined to exist!

- Someone may claim to be able to imagine invisible, intangible fairies at the bottom of the garden smiling when the sun comes out.

  But such imaginings are worthless because since the fairies have no effects on anything, the imagined hypothesis can take infinitely many forms that are indistinguishable: two fairies, three fairies, four fairies; talking English, or German, or Swahili, or ...
  See `http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#pach`

- Someone may claim to be able to imagine finding a set of objects that produce different numbers when counted in different orders, though the objects do not change.

  People can imagine many things that are later proved mathematically impossible.

Representability does not imply possibility

Arguments of the form

> "I can imagine something that has XYZ without phenomenal consciousness, therefore it is possible for XYZ to exist without phenomenal consciousness",

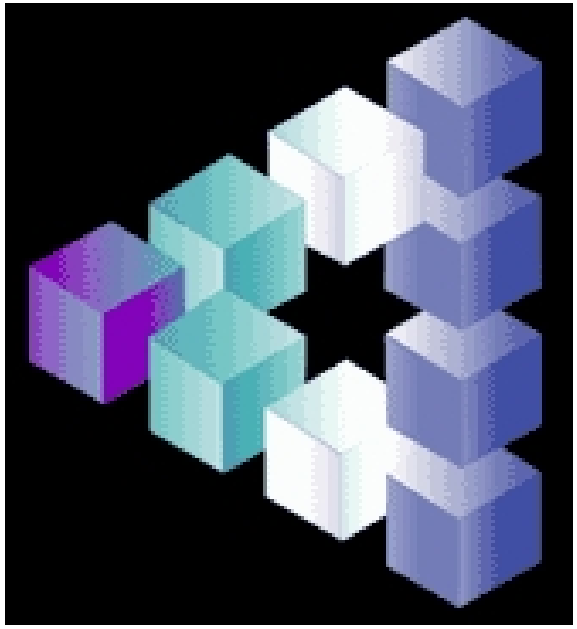
are reminiscent of this claim in Wittgenstein's *Tractatus* (1922):

> **3.0321** "We could present spatially an atomic fact which contradicted the laws of physics, but not one which contradicted the laws of geometry."

Well, here is a spatial representation of a round blue square, seen edge-on:

[———————————] demonstrating that imaginability, does not imply possibility.

Moreover, in the picture by Oscar Reutersvärd, shown earlier, there is a fairly rich and detailed, but nevertheless impossible configuration of cubes. So:



Even quite detailed description, depiction, or representation, of some alleged possibility may conceal a deep contradiction or incoherence.

Untutored philosophical imaginings or intuitions about possible kinds of information-processing systems, or minds, should be treated with deep scepticism, like untutored mathematical imaginings, e.g. imagining discovering the largest prime number.

Someone without experience of designing working information systems may be unable to imagine sophisticated working systems in detail, and may be deceived by assumed implications of superficial (amateur) imaginings – e.g. imagined zombies.

The problem of identity of spatial regions

Someone who is new to philosophy of space and time, and has not studied much physics might ask:

Where is the volume of space that was occupied by my nose exactly ten minutes ago?

- If you are sitting in your office working at your desk and have not moved much in the last ten minutes, the answer could be: “It is now still where it was ten minutes ago”.
- But suppose you were sitting in a railway carriage reading a book for the last ten minutes: you could give different answers, e.g. the location in the carriage, or the portion of the railway track you were close to ten minutes ago.
- Even while at your desk you were on a planet both rotating and hurtling through space, so perhaps your nose is thousands of miles away from its previous location?
- Such examples may lead you to conclude that volumes of space do not have an identity except relative to a reference frame provided by a set of space occupants.

  (As Leibniz did – though Newton disagreed with him.)
- But a defender of absolute space (e.g. Newton) might ask: Where is that volume of space **in itself**, not just in relation to you or me or a carriage or even the earth?
- Such a philosopher might even claim to be able to imagine the volume of space continuing to exist, so it must be somewhere.
- But if spatial location is necessarily relative to space-occupants, then the question where the volume is in itself is incoherent: there is no answer.
- We could label that incoherent problem: the hard problem of spatial re-identification!

Incoherent hard problems

Wondering about the colours of invisible fairies in the garden, the present location of a previously occupied volume of space, the direction of motion of the whole universe, about the time at the centre of the earth, or whether some numbers get hungry at night are all fairly obviously concerns about incoherent questions, and I am suggesting that some problems about consciousness (but not all) are similarly disguised nonsense.

Examples of disguised nonsense questions are:

- whether you have qualia like mine not only in their functional relations but in their intrinsic non-functional qualities (compare the problem of absolute identity of places)
- whether the colour and shape qualia you had yesterday are the same as the ones you have today in the same situations
- how brains can produce functionless phenomenal consciousness

One benefit of developing sharp philosophical intuitions (not part of general education, alas) is learning to detect disguised nonsense: like the “hard” problem of consciousness.

- Of course that still leaves the deep and difficult problems of explaining the widely agreed and easily verified empirical phenomena, of the sorts described earlier, including many phenomena that are only privately accessible.
- Part of the problem is to explain why they are only privately accessible: and the answer proposed here is: because the contents of internal data-structures of complex virtual machines are not accessible from behaviours and not measurable via physical measuring devices (e.g. brain scanners).
- So they need their own modes of investigation, which includes forming rich and detailed conjectures about how the systems work, and testing those conjectures in terms of their detailed consequences, including consequences observable in working models.

Argument from infallibility

Some philosophers have mistakenly been impressed by the infallibility of certain kinds of introspection.

"We are infallible about what we experience."

"We have 'direct access' to our own states of consciousness so we cannot be mistaken about them."

Descartes: "I can doubt everything but not that I am doubting."

"I can be wrong about whether there is a dagger before me,
but not about whether there **seems to me** to be a dagger before me."

But this is exactly like a certain sort of infallibility of every measuring device –

it cannot be mistaken about what it reports!

A voltmeter can be wrong about what the voltage is, if it is faulty, but it cannot be wrong about what it reports the voltage to be.

In the same way, we can be mistaken about what we have seen, but not about what we seem to have seen.

However, nothing of any interest follows from this tautology.

In particular, it does not rule out being mistaken about what we have perceived.

Elaboration/Digression:

The next three slides present an experiment that demonstrates that people sometimes can be mistaken about what they have experienced, though the experiment does not work for everyone.

Experiment demonstrating fallibility about what you see

Experiment (slide 1)

The experiment is presented on the next slide, where a sentence is presented inside a frame.

Your task is to look carefully at what is in the frame and decide whether there is anything wrong with the sentence.

Some people think something is wrong, while others don't see anything wrong.

When you have decided please go to the following slide, where you will be asked some questions.

The experiment does not work for everyone.

In particular, it will not work for you if you have seen this sort of display previously.

Now look at the next slide. If you are sharing this display with someone else, neither of you should say anything about what is in the frame until you are instructed to communicate in the third slide.

Note: this experiment is also available on its own on the internet, here:

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

This may not work for everyone.

Experiment (slide 2)

If you are doing this test with someone else, do not say anything out loud about what you see here:

Trespassers will

will be prosecuted

Is there anything wrong with the sentence in the frame?

Look carefully at the sentence in the frame, and decide whether there is anything wrong with it, and when you have decided (either way) continue to the next slide, without looking back at this one.

Do not say anything to anyone else doing this experiment at the same time, as you may spoil it for them.

Post-test debriefing

Experiment (slide 3)

If you saw something wrong with the sentence in the frame while you were looking at it then this experiment has not worked for you.

If you saw **nothing** wrong please try to answer these two questions **without** looking back at the previous slide:

- **How many words were in the frame?**
- **Where was the word "will" ?**

If thinking about one or other of those questions (without looking back) makes you realise that what you saw had an error that you did not previously notice, then the experiment has worked for you.

If the experiment worked for you, then you were able, simply by asking yourself a new question, to notice an aspect of your experience that had earlier escaped your attention.

That demonstrates that you can be mistaken about the contents of your own consciousness when looking at the original display because the information you got from the display included something you thought was not present: you thought nothing was wrong. (This happens often during proof-reading of text.)

The information about the duplicated word must have been in your mind if you could answer either of the above questions without looking back at the display.
(Now you may talk, and look back at the previous slide.)

Mythical Infallibility

The alleged infallibility of knowledge about the contents of one's consciousness is supposed to be evidence that people have direct understanding of the nature of phenomenal consciousness.

Arguments based on that kind of infallibility are flawed in two ways:

- As we have seen, there are aspects of their own contents of consciousness about which people can be mistaken (e.g. when failing to spot errors in texts they are reading).
- Insofar as there are real examples of infallibility, they are the tautological consequences of trivial truths similar to
  - I cannot be wrong about how things seem to me.
  - A voltmeter cannot be wrong about the voltage it has measured.

Claims that such infallibility demonstrates the undeniable existence of something (seemings, qualia, ...) are therefore not to be taken seriously.

Summary of the last few slides

A: Arguments of the form "I can imagine something with mechanisms of type X, but lacking this sort of experience" are not to be taken seriously as arguments against attempts to explain having experiences in terms of mechanisms of type X, since humans, including sophisticated philosophers, can imagine things that are impossible without realising that that is what they are doing.

Sometimes that can be because they imagine the wrong (over-simple) things of type X, e.g. if they don't know enough about varieties of information processing systems.

B: Claims to know about the nature of consciousness on the basis of infallible types of introspection are not to be taken seriously, since

(a) introspection can be fallible

(b) the infallible aspect is tautological and similar to infallibility found in all representational devices: they can't be wrong about what they are representing, though what they represent can be wrong.

The upshot is that common philosophical arguments intended to refute the claim that minds are information-processing systems can be discounted.

But we still lack a deep theory about relevant types of information processing systems: including how they evolved and how they work. But we have made some progress.

We need to understand minds with architectures

The preceding experiment gives us clues about the **architecture** of a human mind, and some aspects of the structure of visual experience:

- A mind is not a unitary entity whose contents of consciousness at a time are sharply defined.
- Rather a mind is a complex information-processing system with an architecture that involves different subsystems working in parallel, and the subsystems may acquire, use and store different information from the same sensory input. (Different kinds of minds have different interacting subsystems.)
- In particular, the subsystem that is engaged in communication with someone else (e.g. trying to answer the question I posed for you earlier about the contents of the box) may fail to detect part of what is perceived, even though another subsystem has all that information and is capable of being interrogated by the communication subsystem even after the original physical source of the information has been removed.
- What is perceived (the contents of visual experience) can have different layers of information, including a layer concerned with the distribution of colour and texture in the visual field, a layer concerned with "chunks" in the field, layers concerned with more abstract entities, such as phrases and sentences, and many others ...

  See Chapter 9 of *The Computer Revolution in Philosophy* (1978)
  `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/chap9.html`

Varieties of having and using information

It is often assumed that if something has and uses information it also has the information that it has the information: but this is a fallacy.

- For organisms, much (often all) of the information used is transient control information.

  If an organism senses a noxious substance and shrinks away from it, the information that led to the action could be transient: there need be no subsequent record of it.

- Being able to use information for control purposes is one thing: retaining the information for future use is another.

  It requires a more complex information-processing architecture, allowing previously used information to be stored and accessed: we don't notice that when we experience or remember change.

- One important reason for storing information after use is that it can make change detection possible: otherwise an organism merely has at any instant information about that instant.
- The ability to detect change (and processes, flows, speeds at which things happen) requires more than just storage after use

  the architecture needs to include mechanisms that compare stored and new information: to produce information about the change (e.g. this is crucial for homeostatic control mechanisms).

- Information about change also could be either transient, or recorded for future use.
- People are surprised at change blindness: **instead they should be puzzled by change-detection, and by detection of no-change!**
- So detecting change or persistence, using change, detecting change detection, storing information about what does and does not change, all require additional information processing mechanisms beyond merely having and using information.

  Similar comments apply to detection of **spatial** differences, curvature, gradients, boundaries, etc.

The mythical "it"

We have a noun "consciousness" and we assume there must be something it refers to: forgetting that it is derived from an adjective "conscious", whose primary use is before a prepositional phrase "of X".

Many spurious problems arise from such reifications, e.g. asking what the noun "emotion" refers to instead of analysing the adjectives "emotional", and "moved", and related adjectives, "afraid", "despondent", "grateful", "infatuated", "devastated", etc.

The use of the noun "consciousness" suggests there is one thing referred to and invites questions about "it", like

What are its functions?
Which animals have it?
At what stage does a foetus have it?
Which bits of brain produce it?
How did it evolve?
Can machines have it?
What causes it?
How does the brain produce it?

Attending to the use of the adjective with "of ...", reveals that there is no one thing.

There is no "it" about which we can ask such questions, because there are huge differences between being conscious (or not) of a toothache, of how late it is, of the changing weather, of the day of the week, of being puzzled, of being unable to remember a name, of being unpopular, of being in favour of capital punishment, of a likely collision, and many more.
"Conscious" is a highly polymorphic concept. (Familiar in computing.)

Adjectives like "efficient", and "useful" are also polymorphic, in similar ways.

There is no one thing that being efficient, or useful, or expensive or dangerous is: in each case, what the adjective indicates has to be specified in the context, e.g. being an efficient (or dangerous) lawn-mower or microwave oven, proof strategy, government department, or murderer.

Semantic polymorphism: multiple "ITs"

It is well known that the meaning of the adjective "tall" does not specify a minimum height required for being tall: some additional noun or noun phrase is required to specify a comparison class, e.g.

A tall squirrel, a tall giraffe, a tall building for a small rural town, a tall flea?

These are cases of "parametric polymorphism": an extra parameter is required to specify which of several interpretations the word has.

The same applies to "A is conscious of X": What A and X are make a big difference.

There are many consequences

- Different animals can be conscious of different things, and with different consequences

  and in different ways (e.g. echolocation vs using feelers vs using vision, etc.)

- So there are different varieties of consciousness that evolved at different times in different species – these are not minor differences of degree.
- Different information-processing mechanisms and architectures are required
  - in some cases the mechanisms require introspective mechanisms (e.g. being conscious of feeling annoyed or surprised), but it's likely that not all animals can do that;
  - other kinds of consciousness (e.g. being conscious of an obstacle to be avoided or of something large approaching quickly) seem to be in many more species;
  - some forms of human consciousness are not possible for infants or toddlers: the information processing architecture is still too primitive (e.g. being conscious of finding algebra difficult);
  - some are only possible in particular cultures, e.g. being conscious of having begun a new millenium, was impossible for our early homo sapiens ancestors. (Why?)

How did consciousness evolve? How did what? evolve?

Can we draw a line between things with and without consciousness?

- Some people (e.g. Susan Greenfield) think it's all just a matter of degree: evolution produced more and more complex kinds of consciousness.

  BUT evolution cannot produce continuous change –
  so we need to study many small discontinuities rather than one big one or only smooth transitions.

- One way to think about this:
  - instead of trying to draw a single boundary, regard anything that acquires and uses information (i.e. all living things) as in some way conscious
  - but they are conscious of (acquire information about) different things; and their consciousness has different functions and effects. There are differences in:
    - what information is acquired
    - how it is acquired
    - how it is processed
    - how it is used, and whether different sub-systems use it in different ways
    - how it is stored
    - how it is encoded/represented, and whether it is stored in different ways in one individual,
    - how much of it is represented in features of the environment rather than entirely in the user.
  - we can then try to understand the many different problems posed to different species, and the many different solutions produced by various combinations of biological evolution, cultural evolution, individual development and learning.... a survey of types of information-processing system
  - Instead of differences of degree we have differences of kind – not necessarily linearly ordered.

- This requires good conceptual tools for thinking about such systems, how they work, and how they differ: What tools? – tools for designing testable working systems!

The ordinary notion vs the scientific notion

Scientists and philosophers who don't understand the diversity associated with the polymorphism of "consciousness" as ordinarily understood, are tempted to formulate pointless, unanswerable questions about "it".

The ordinary notion of consciousness (being conscious of XX) is highly polymorphic.

There are major differences in varieties of consciousness (being conscious of something) that depend on what XX is: different XXs require different information-processing competences and architectures.

Some of those are products of recent evolution, or cultural development, or individual learning, while others are based on evolutionarily old mechanisms. (See previous slides.)

People who don't notice this diversity/polymorphism talk about consciousness as one "it":

- How did **it** evolve? or When did **it** evolve?
- What are **its** functions?
- Which things have **it**? (At what stage is **it** in a foetus?)
- What are **its** neural correlates?

Instead of assuming that there's one thing all these questions refer to, we need to allow that there are different things and each question has a variety of answers.

So searching for neural correlates of "it" is pointless: though there may be neural correlates of some of "them". (As explained later, some may have no neural correlates.)

However it takes hard analytical research (requiring philosophical sophistication) to identify the diversity of "them", so as to clarify what the various research problems are.

None of this matters for ordinary usage

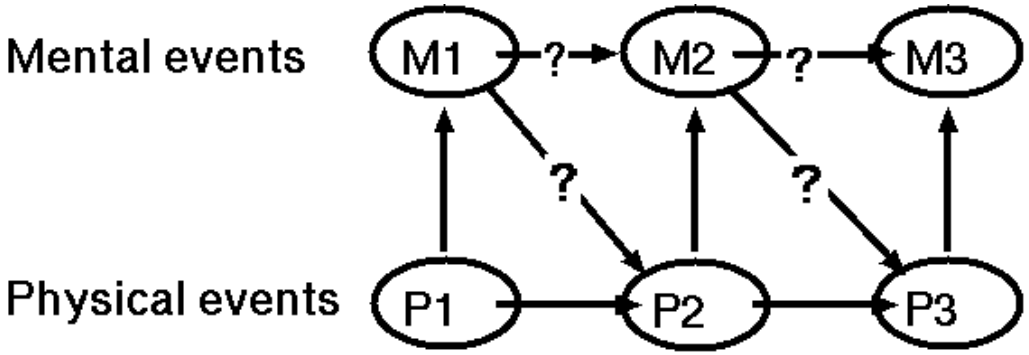
The problems arise only when philosophers and scientists make false assumptions about what "consciousness" and related words mean, since in our ordinary usage we take the polymorphism in our stride: we have learnt to make use of context in interpreting uses of many polymorphic concepts, including "tall", "efficient", "useful", "dangerous", "prevention", "aware", "attend", and "conscious".

- These are examples of the extreme complexity of unconscious processing that goes on when we learn and use language: we acquire and use very many complex syntactic rules that we are totally unaware of, some of which are quite hard to discover, as linguists have found.
- In addition to ordinary informal usage, as in "I have been conscious of your hostility to me", there are many medical uses of the adjective "conscious" and noun "consciousness" which do not make any explicit claims about what consciousness is, e.g. when a doctor asks "when did the patient regain consciousness?"
- It is only when consciousness itself becomes the subject of research, as if it were a well-defined unitary topic, that the confusions I am criticising tend to arise.
- If consciousness is not a unitary phenomenon with specific neural correlates, we can ask how the variety of non-physical phenomena found in minds can occur as a result of the operation of different kinds of physical machinery, some produced at different stages of biological evolution, as found in brains, sensors and effectors.

Causation is at the heart of the problem

Is this an acceptable picture of causal relations between brain and mind?

If M1, M2, M3 fit the definition of "phenomenal consciousness" then all the arrows with "?" must be removed.
P-C is defined to be epiphenomenal – a side effect of physical processes, but lacking any functional role, which means P-C has no causal powers.



However, if M1, M2, M3 fit the definition of A-C, they can have cognitive functions and causal powers.

But how can they influence physical events if all of P1, P2, P3 are determined by previous physical events?

Worries about this leads some people to propose gaps in physical causation (e.g. quantum indeterminacy) to allow mental causation to play a role.

We don't need any gaps in physical causation.

In the last half century we have learnt how to build active virtual machines (VMs) that run in parallel with physical machinery on which they depend for their existence, and with the ability to control behaviours of those physical machines.

Objections to this claim are based on an inadequate understanding of causation, as if it were some sort of fluid flowing through time.

Instead we need to understand causation in terms of networks of structural relations that constrain what can happen, including relations and constraints involving entities and processes in running virtual machines, e.g. operating systems, spelling checkers, ...

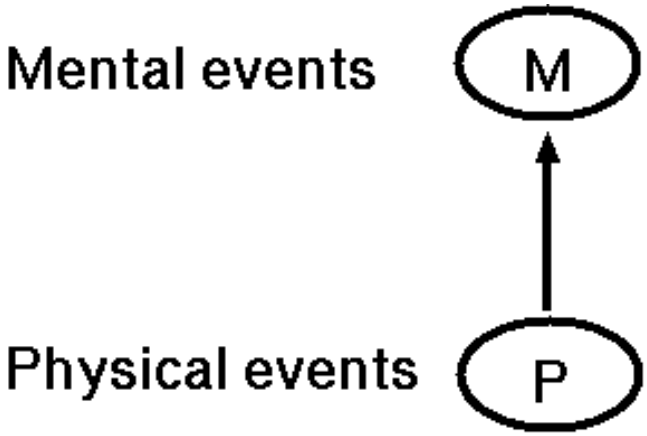
Towards a meta-theory of information-processing systems

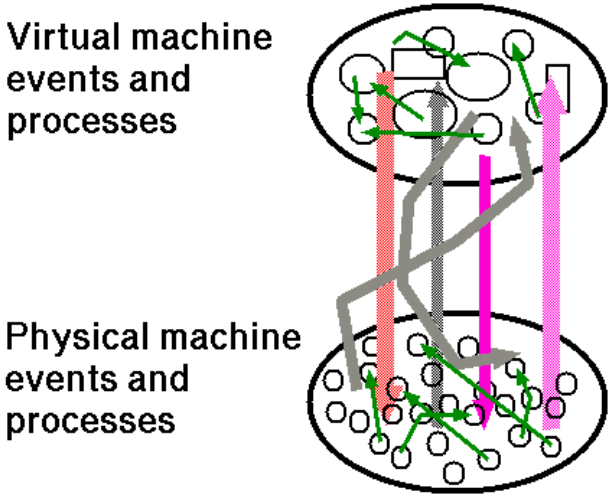
- The most advanced such theories come from computer science/software engineering.
- They have led to profound technological advances in information processing machinery.
- But most computer scientists ignore the biological information processing systems that preceded the development of electronic computers.
- So their theories are not general enough
- Some very important. complex and technical, ideas have been developed since the 1950s and we can use them, and extend them in trying to understand products of biological evolution.
- Among the most important is the idea of a running virtual machine (RVM), not to be confused with the abstract mathematical structure defining a type of VM.

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

The 20th C Philosophical breakthrough: Virtual machinery

Brief introduction to the technology of virtual machinery (not virtual reality) developed over the last six decades and its philosophical relevance:
Processes and events in running virtual machines can be causes and effects, despite being implemented in deterministic physical mechanisms.

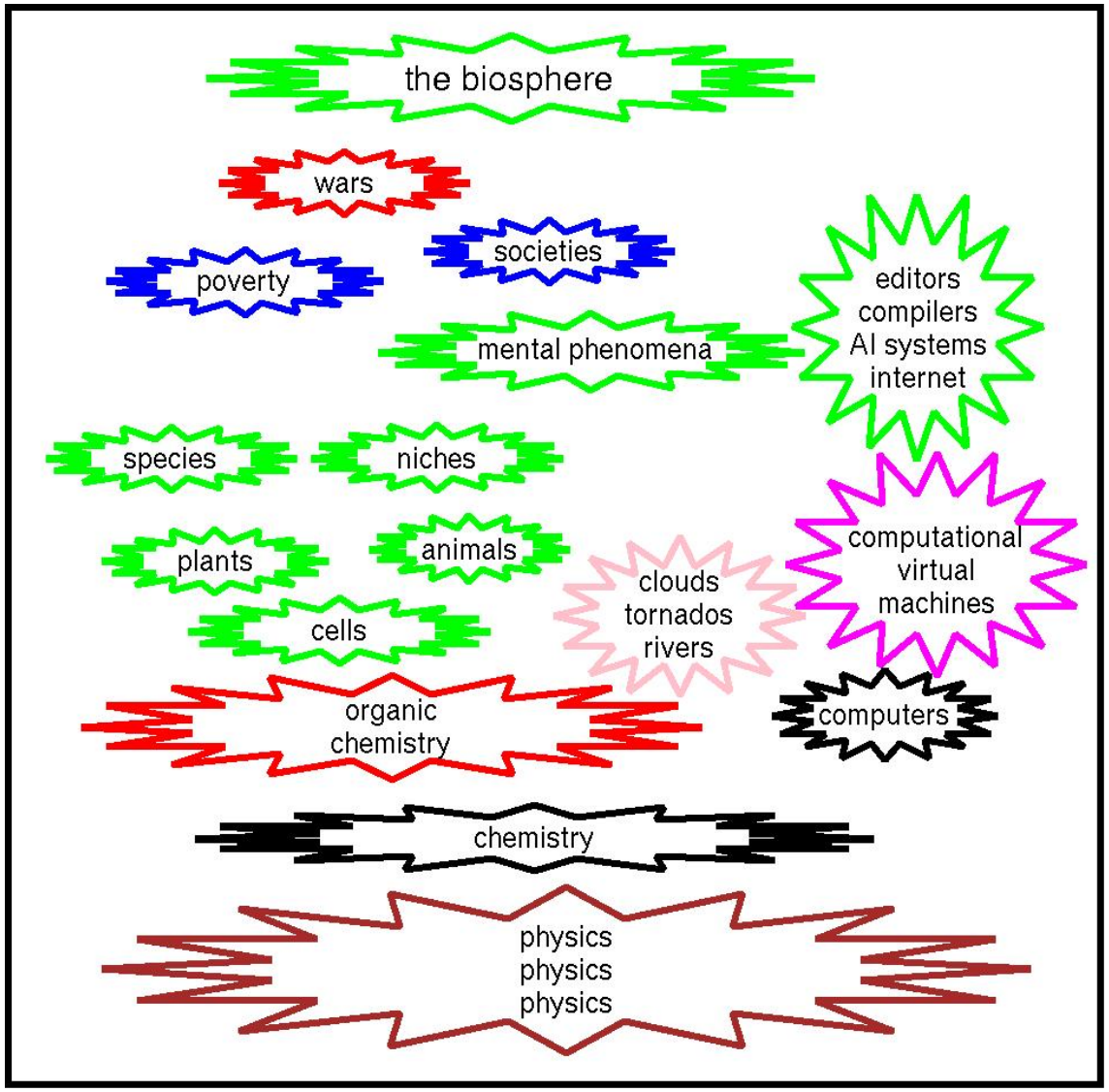




Instead of the picture on the left, which implies that there is only one-way causation from physical to mental, we need to understand how running virtual machinery can co-exist with, and influence, underlying physical machinery, which it helps to control, even though the virtual machinery is all fully implemented in the physical machinery.

The virtual machinery (e.g. a chess program, or email program) chunks processes at a level of abstraction that cannot be defined in the language of physics, as explained in this presentation: `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#mos09`

Virtual machines are everywhere

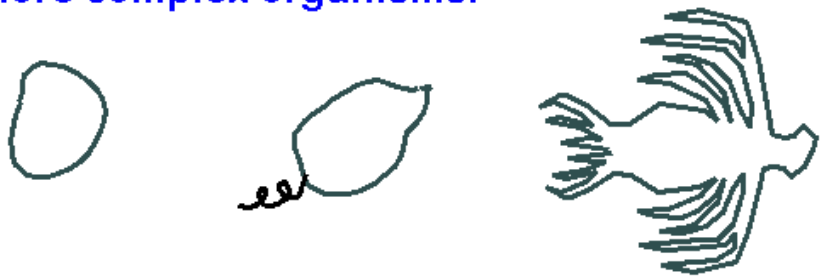


How many levels of (relatively) virtual machinery does physics itself require?

A brief history of evolution!

Once upon a time there were only inorganic things: atoms, molecules, rocks, planets, stars, etc. These merely reacted to *resultants* of all the physical forces acting on them.

Later, there were simple organisms. And then more and more complex organisms.



These organisms had the ability to reproduce. More interesting was their ability to *initiate* action, and to *select* responses, instead of simply being pushed around by physical forces acting on them.

That achievement required the ability to acquire, process, and use *information*.

Growing environmental demands, and increasing complexity of the organisms themselves, led to many important changes in the requirements for information-processing mechanisms and architectures, forms of representation, and ontologies.

Eventually the complexity of the control problems demanded use of virtual machinery running on the biological physical machinery.

(Physical reconfiguration (e.g. neural re-growth) is much too slow for animals in dynamic environments!)

Mechanisms required to support running VMs

We don't know exactly what problems evolution faced, and what solutions it came up with, and what mechanisms it used, in creating virtual machinery running in animal bodies, to control increasingly complex biological organisms (and societies), but perhaps we can learn from some of the developments supporting increasing use of increasingly sophisticated VMs in artificial systems over the last half century (a small sample follows):

- The move from bit-level control to control of and by more complex and abstract patterns.
- The move from machine-level instructions to higher level languages (using compilers that ballistically translate to machine code and especially interpreters that "translate" dynamically, informed by context).
- Memory management systems make physical memory reference context-dependent. (
- Virtual memory (paging and cacheing) and garbage collection switch virtual memory contents between faster and slower core memories and backing store, and between different parts of core memory: constantly changing PM/VM mappings. (These support multiple uses of limited resources.)
- Networked file systems change apparent physical locations of files.
- Device interfaces translate physical signals into "standard" VM signals and vice versa.
- Devices can themselves run virtual machines with buffers, memories, learning capabilities...
- Device drivers (software) handle mappings between higher level and lower level VMs – and allow devices to be shared between VMs (e.g. interfaces to printers, cameras, network devices).
- Context-dependent exception and interrupt handlers distribute causal powers over more functions.
- Non-active processes persist in memory and can have effects on running processes through shared structures. (It's a myth that single-cpu machines cannot support true parallelism.)
- Multi-cpu systems with relocatable VMs allow VM/PM mappings to be optimised dynamically.
- Multiplicity of concurrent functions continually grows – especially on networked machines.
- Over time, control functions increasingly use monitoring and control **of VM states and processes.**

Different requirements for virtual machinery

The different engineering developments supporting new kinds of virtual machinery helped to solve different sorts of problems. E.g.

- Sharing a limited physical device between different users efficiently.
- Optimising allocation of devices of different speeds between sub-tasks.
- Setting interface standards so that suppliers could produce competing solutions.
- Allowing re-use of design solutions in new contexts.
- Simplifying large scale design tasks by allowing components to "understand" more complex instructions (telling them what to do, leaving them to work out how to do it).
- Specifying abstract kinds of functionality that could be instantiated in different ways in different contexts (polymorphism).
- Improving reliability of systems using unreliable components.
- Allowing information transfer/information sharing to be done without constant translation between formats.
- Simplifying tasks not only for human designers but also for self-monitoring, self-modulating, self-extending systems and sub-systems.

These are solutions to problems that are inherent in the construction and improvement of complex functioning systems: they hare not restricted to artificial systems, or systems built from transistors, or ...

Conjecture: Similar problems were encountered in biological evolution (probably many more problems) and some of the solutions developed were similar, while some were a lot more sophisticated than solutions human engineers have found so far.

Virtual machinery and causation

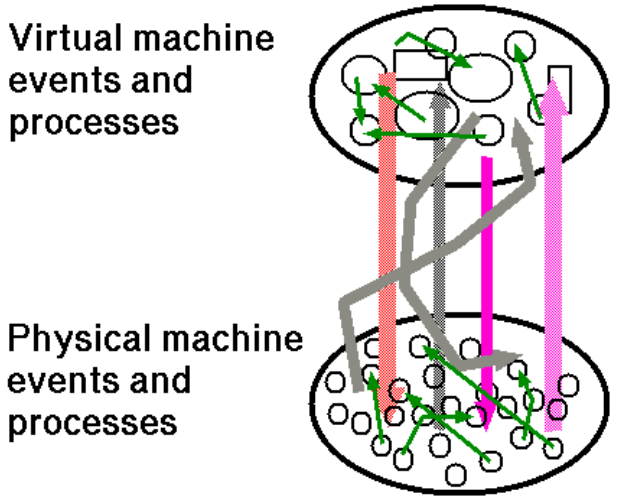
Virtual machinery works because "high level" events in a VM can control physical machinery.

Accordingly, bugs in the design of virtual machines can lead to disasters, even though there's nothing wrong with the hardware.

As stated previously

> Processes and events in running virtual machines can be causes and effects, despite being implemented in deterministic physical mechanisms.

Engineers (but not yet philosophers, psychologists, neuroscientists?) now understand how running virtual machinery can co-exist with, and influence, underlying physical machinery, which it helps to control, even though the virtual machinery is all fully implemented in the physical machinery.



This works because

- We have learnt how to set up physical mechanisms that enforce constraints between abstract patterns (in contrast with mechanisms that enforce constraints between physical or geometric relations).
- Chains of such constraints can have complex indirect effects linking different patterns.
- Some interactions involve not only causation but also meaning: patterns are interpreted as including descriptive information (e.g. testable conditions) and control information (e.g. specifying what to do).

  See "What enables a machine to understand?" (IJCAI 1985)
  `http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#4`

Could biological evolution have solved similar problems?

Physical $\Longleftrightarrow$ virtual interfaces at different levels

Starting with simple physical devices implementing interacting discrete patterns, we have built layers of interacting patterns of ever increasing spatial and temporal complexity, with more and more varied functionality.

- Physical devices can constrain continuously varying states so as to allow only a small number of discrete stable states (e.g. only two)
  (e.g. using mechanical ratchets, electronic valves (tubes), aligned magnetic molecules, transistors etc.)
- Networks of such devices can constrain relationships between discrete patterns.
  E.g. the ABCD/XY example: a constraint can ensure that if devices A and B are in states X and Y respectively then devices C and D will be in states Y and X (with or without other constraints).
  So, a device network can rule out some physically possible combinations of states of components, and a new pattern in part of the network will cause pattern-changes elsewhere via the constraints.
  Compare: one end of a rigid lever moving down or up causes the other end to be moving up or down.
- Such networks can form dynamical systems with limited possible trajectories, constraining both the possible patterns and the possible sequences of patterns.
- A network of internal devices can link external interfaces (input and output devices)
  thereby limiting the relationships that can exist between patterns of inputs and patterns of outputs, and also limiting possible sequences of input-output patterns.
- Patterns in one part of the system can have meaning for another part, e.g.
  - constraining behaviour (e.g. where the pattern expresses a program or ruleset) or
  - describing something (e.g. where the pattern represents a testable condition)
- Such patterns and uses of such patterns in interacting computing systems may result from design (e.g. programming) or from self-organising (learning, evolving) systems.
- Some useful patterns need not be describable in the language of physics.

Try playing with interacting abstract patterns

There are movies showing some videos of some interacting virtual machines implemented in the (freely available) Poplog SimAgent Toolkit here:

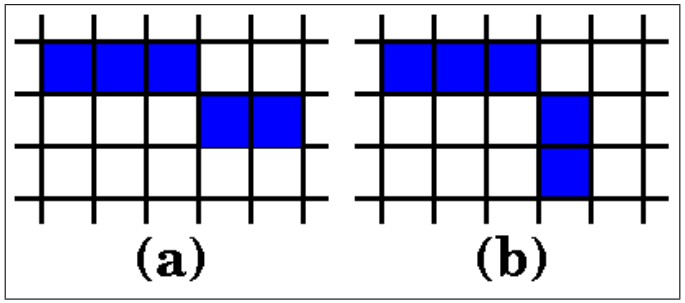
`http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html`

John Conway's "Game of Life" provides many examples of interacting abstract patterns.

An excellent online version can be run in a (java-enabled) web browser here:
`http://www.bitstorm.org/gameoflife/` (It allows mouse interaction with a running process.)

Anyone can play with interacting patterns by clicking on squares to make an initial pattern or set of patterns (or selecting a pre-built pattern) from the menu labelled 'Clear', and then stepping through the interactions by repeatedly pressing 'Next', or launching the process by pressing 'Start'.

For example, a starting pattern of five squares like the pattern (a) very quickly fizzles out, whereas moving one of the blue squares to form the starting pattern (b) on the right produces a totally different result: once started the display continues changing hundreds of times without repeating itself until it eventually (after more than 1100 changes – if the board is big enough) gets stuck with a collection of static or repeating small patterns.



Such demonstrations illustrate ways in which abstract patterns can be made to interact causally (parts of patterns cause other parts to change), by harnessing physical machinery to operate according to a set of constraints (Conway's four rules).

The constraints can be implemented via mechanical linkages, or via electronic circuits, or via a computer program that forces the computer hardware to behave as if wired to obey the rules: as a result a group of squares changing state can cause a new pattern to come into existence.

Which patterns interact?

When Conway's machine runs, do patterns on the screen interact causally? NO!

The screen merely displays what is going on in a virtual machine running in the computer.

In the running VM the abstract data-structures in the 2-D grid **in the VM** interact.

Those changing VM structures are represented by changing squares on the screen merely to inform us what is happening inside.

- In a different physical design the screen could be part of the implementation of the VM.
  Most Conway implementations would go on running even if the screen were disconnected: they would merely cease to display anything. That is how some people understand contents of consciousness – i.e. as causally/functionally ineffective (e.g. P-C as defined by Block).
  (Perhaps they think zombies are possible because they imagine turning off an internal screen.)
- If the grid information used in applying the rules were stored in devices implementing the pixels of the screen display, then we could say that the cells in the screen interact – but not their visual effects.
- However, if you are inspired, by what you see on the screen, to interact by clicking in the grid to change a pattern, altering future behaviour, then the visible screen patterns are causally effective in changing how the process develops, and your eyes and brain become part of the physical implementation of a very sophisticated distributed virtual machine!

The causation observed in a Conway machine is normally only in the (invisible) virtual machine that causes what is on the screen to change: the screen display is (relatively) epiphenomenal, like changing expressions on people's faces, when they are alone!

For more examples and references see `http://en.wikipedia.org/wiki/Conway%27s_Game_of_Life`

More complex interacting patterns

If a run of the machine starts with two separated patterns, then for a time each may change internally without influencing the other, even if each pattern has many internal interactions.

But that can change after a while.

- The distances between cells influenced by the two patterns may gradually be reduced either because the patterns grow, or because they move (e.g. so-called “gliders” in Conway’s machine).
- The results of the interaction may feed back into both patterns making them behave thereafter very differently from how they would have behaved on their own.
- The form of interaction can depend not only on the starting shape of each pattern, but exactly how far apart they are initially (and whether they “bump into” the grid frame).
- In other cases they may interact via a communicating bridge pattern.
- In all cases the interactions are caused by pattern structures and pattern relationships insofar as the same patterns placed in different starting locations of the Conway machine will always produce the same results: so the precise implementation mechanism is less important than the patterns and their rules of behaviour.

It is also possible to produce variants of the Conway machine whose rules are probabilistic (stochastic) rather than deterministic.

The form of causation in such machines is then probabilistic, except when pattern structures constrain possible outcomes despite the probabilistic rules.

Reactive vs deliberative interacting patterns

A Conway machine uses real or simulated concurrency: behaviour of each square depends only on the previous states of its eight neighbours and nothing else.

On a computer the concurrency is achieved by time-sharing, but it is still real concurrency.

Consider what happens when two virtual machines running on a computer compete in a chess game, sharing a virtual board, and interacting through moves on the board, each can sense or alter the state of any part of the (simulated) chess board.

- In general, programs on a computer are not restricted to local interactions.
- In some cases, the interacting processes are purely reactive: on every cycle every square immediately reacts to the previous pattern formed by its neighbours.
- If two instances of a chess program (or instances of different chess programs) interact by playing chess in the same computer, their behaviour is typically no longer purely reactive. Good ones will often have to search among possible sequences of future moves to find a good next move – and only then actually move.

  In addition, one or both of the chess virtual machines may do some searching in advance while waiting for the opponent's next move.
- Then each instance is a VM with its own internal states and processes interacting richly, and a less rich interaction with the other VM is mediated by changes in the shared board state (represented by an abstract data-structure).

  For more on varieties of deliberation see:
  `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604`

Intentionality in a virtual machine

A running chess program (a VM) takes in information about the state of the board after the opponent moves, and builds or modifies internal structures that it uses to represent the board and the chess pieces on it, and their relationships, including threats, opportunities, possible traps, etc.

- In particular it uses those representations in attempting to achieve its goals.

  So, unlike the interacting Conway patterns mentioned earlier, some of the patterns in the chess virtual machine are treated by the machine as representations, that refer to something.

- During deliberation, some created patterns will be treated as referring to non-existent but possible future board states, and as options for moves in those states.

  They are treated that way insofar as they are used in considering and evaluating possible future move sequences in order to choose a move which will either avoid defeat (if there is a threat) or which has a chance of leading to victory (check-mate against the opponent).

- In this case the chess VM, unlike the simplest interacting Conway patterns, exhibits intentionality: the ability to refer. (NB. The programmer need not know about the details.)

  Since the Conway mechanism is capable of implementing arbitrary Turing machines, it could in principle implement two interacting chess virtual machines, so there could be intentionality in virtual machines running on a Conway machine – probably requiring a very big fairly slow machine.

- The intentionality of chess VMs is relatively simple because they have relatively few types of goal, relatively few preferences, and their options for perceiving and acting are limited by being constrained to play chess:

  For a human-like, or chimp-like, robot the possibilities would be much richer, and a far more complex architecture would be required. See
  `http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200307`

Adding meta-semantic competence

If a virtual machine playing chess not only thinks about possible board states and possible moves, and winning, moving, threats, traps, etc. but also thinks about what the opponent might be thinking, then that requires meta-semantic competences: the ability to represent things that themselves represent and use information.

- It is very likely that biological evolution produced meta-semantic competences in some organisms other than humans because treating other organisms (prey, predators, conspecifics to collaborate with, and offspring as they learn) as mere physical systems, ignoring their information-processing capabilities, will not work well (e.g. hunting intelligent prey, or avoiding intelligent predators).
- Another application for meta-semantic competences is self-monitoring, self evaluation, self-criticism, self-debugging: you can't detect and remedy flaws in your thinking, reasoning, planning, hypotheses etc. if you are not able to represent yourself as an information user.
- It is often assumed that social meta-semantic competences must have evolved first, but that's just an assumption: it is arguable that self-monitoring meta-semantic competences must have evolved first

  e.g. because an individual has relatively direct access to (some of) its own information-processing whereas the specifics of processing in others has to be inferred in a very indirect way (even if evolution produced the tendency to use information about others using information).
  See A. Sloman, 1979, The primacy of non-communicative language,
  `http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#43`

Emerging varieties of functionality

Computer scientists and engineers and AI/Robotics researchers have been learning to add more and more kinds of control, kinds of pattern, and ways of interpreting patterns of varying levels of abstraction.

- A simple machine may repeatedly take in some pattern and output a derived pattern,
  e.g. computing the solution to an arithmetical problem.
- More complex machines can take in a pattern and a derivation-specification (program) and output a derived pattern that depends on both.
- Other machines can continually receive inputs (e.g. from digitised sensors) and continually generate outputs (e.g. to digitally controlled motors).
- More sophisticated machines can
  - solve new problems by searching for new ways of relating inputs to outputs, i.e. learning;
  - interpret some patterns as referring to the contents of the machine (using a somatic ontology) and others to independently existing external entities, events, processes (using an exosomatic ontology)
  - extend their ontologies and theories about the nature and interactions of external entities
  - perform tasks in parallel, coordinating them,
  - monitor and control some of their own operations – even interrupting, modulating, aborting, etc.
    (Including introspecting some of their sensory and other information contents: qualia.)
  - develop meta-semantic ontologies for representing and reasoning about thinking, planning, learning, communicating, motives, preferences, ...
  - acquire their own goals and preferences, extending self-modulation, autonomy, unpredictability, ...
  - develop new architectures which combine multiple concurrently active subsystems.
  - form societies, coalitions, partnerships ... etc.
- Biological evolution did all this and more, long before we started learning how to do it.

Causal networks linking layered patterns

How can events in virtual machines be causes as well as effects, even causing physical changes?

The answer is

through use of mechanisms that allow distinct patterns of states and sequences of patterns to be linked via strong constraints to other patterns of states and sequences of patterns (as in the ABCD/XY example, and the Conway machines, mentioned above). (Some VMs may use probabilistic/stochastic constraints.)

What many people find hard to believe is that this can work for a virtual machine whose internal architecture allows for divisions of functionality corresponding to a host of functional divisions familiar in human minds, including

- interpreting physical structures or abstract patterns as referring to something (intentionality)
- generation of motives,
- selection of motives,
- adoption of plans or actions,
- perceiving things in the environment,
- introspecting perceptual structures and their changes,
- extending ontologies,
- forming generalisations,
- developing explanatory theories,
- making inferences,
- formulating questions,
- and many more.

Biological unknowns: Research needed

Many people now take it for granted that organisms are information-processing systems, but much is still not known, e.g. about the varieties of low level machinery available (at molecular and neuronal mechanisms) and the patterns of organisation for purposes of acquiring and using information and controlling internal functions and external behaviours.

Steve Burbeck's web site raises many of the issues:

"All living organisms, from single cells in pond water to humans, survive by constantly processing information about threats and opportunities in the world around them. For example, single-cell E-coli bacteria have a sophisticated chemical sensor patch on one end that processes several different aspects of its environment and biases its movement toward attractant and away from repellent chemicals. At a cellular level, the information processing machinery of life is a complex network of thousands of genes and gene-expression control pathways that dynamically adapt the cell's function to its environment."

`http://evolutionofcomputing.org/Multicellular/BiologicalInformationProcessing.html`

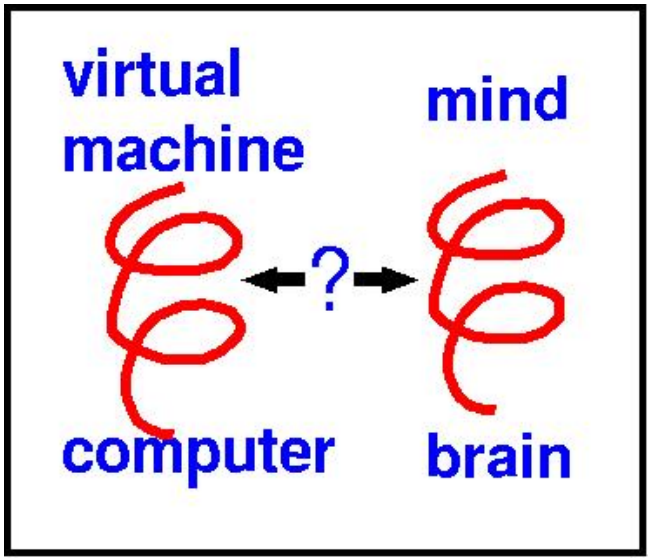
"Nature offers many familiar examples of emergence, and the Internet is creating more.

The following examples of emergent systems in nature illustrate the kinds of feedback between individual elements of natural systems that give rise to surprising ordered behavior. They also illustrate the trade off between the number of elements involved in the emergent system and the complexity of their individual interactions. The more complex the interactions between elements, the fewer elements are needed for a higher-level phenomenon to emerge. ... networks of computers support many sorts of emergent meta-level behavior because computers interact in far more complex ways than air and water molecules or particles of sand ... Some of this emergent behavior is desirable and/or intentional, and some (bugs, computer viruses, dangerous botnets, and cyber-warfare) are not."

`http://evolutionofcomputing.org/Multicellular/Emergence.html`

An old idea.

The idea of the analogy expressed in this diagram is very old, but we are only slowly understanding the variety of phenomena on the left hand side, extending our appreciation of what might be going on on the right.

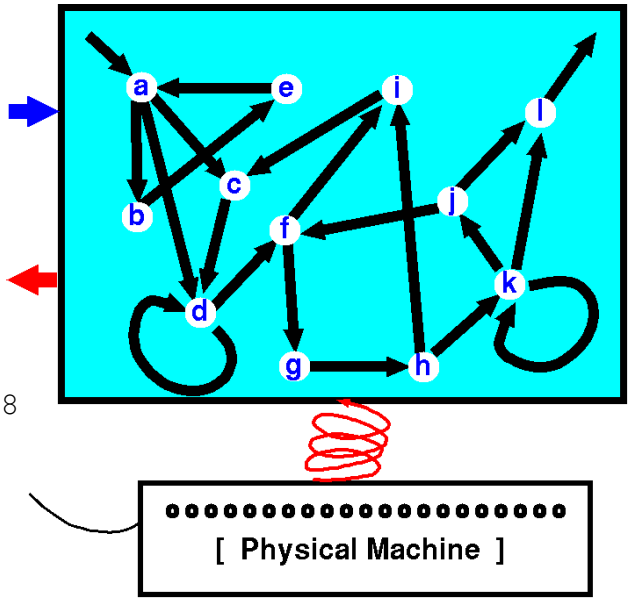


The simple-minded notion that the left hand side involves a program in a computer is seriously deficient, (a) because it ignores the requirement for the program to be running and (b) because we know now that there are far more types of information-processing system than a computer running a single program, as explained in other slides.

Simple-minded views about virtual machines lead to easily refutable computational theories of mind, e.g. the theory that virtual machines are simple finite state machines, as illustrated on the right ("Atomic state functionalism"). See

Ned Block: Functionalism `http://cogprints.org/235/`

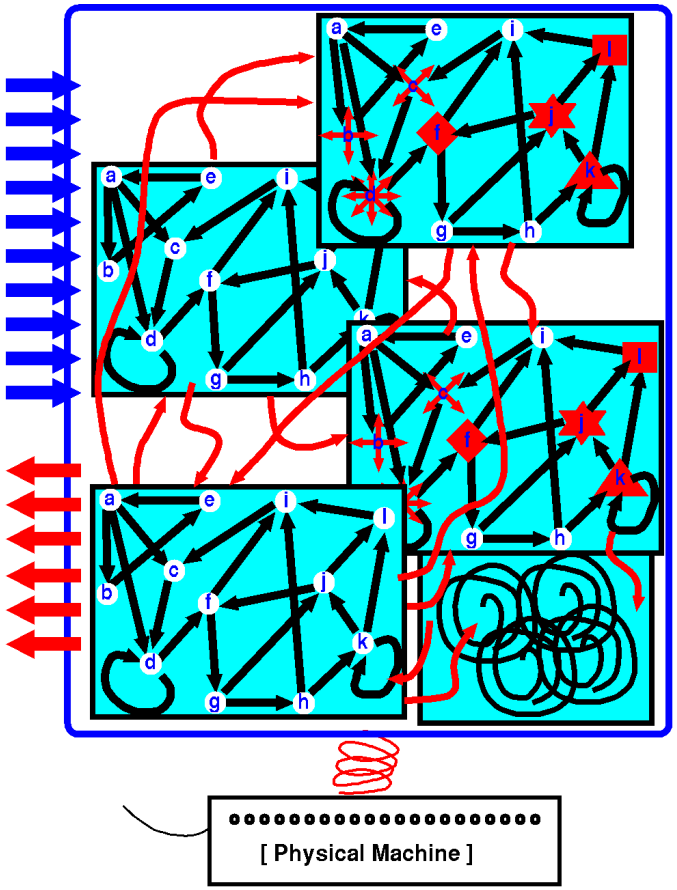
A. Sloman, The mind as a control system, 1993,
`http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#18`

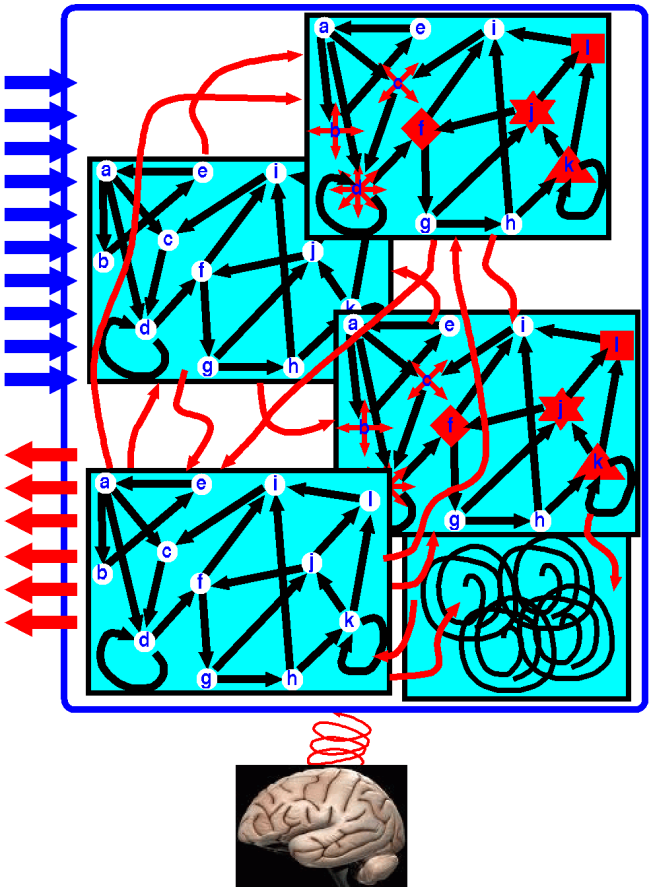


Forget what you have learnt about Turing machines: that's a simple abstraction which is surprisingly useful for theorising about classes of computations – but not so useful for modelling complex multi-component systems interacting asynchronously with a rich and complex environment.

More realistic models

As crudely indicated here we need to allow multiple concurrent inputs (blue) and outputs (red), multiple interacting subsystems, some discrete some continuous, with the ability to spawn new subsystems/processes as needed.





Artificial VM on artificial PM **Biological VM on biological PM**

In both cases there are multiple feedback loops involving the environment.

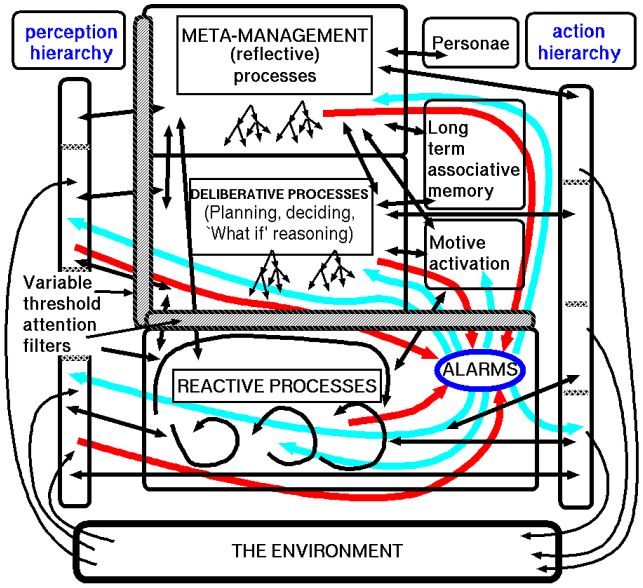
Architectures

We need good conceptual tools for talking about architectures.

The CogAff architecture schema crudely depicted on the left specifies types of components of an architecture.

Particular architectures can be specified by filling in the boxes and indicating connections (information flow and causal influences) between subsystems.

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



A particular way of filling in the boxes has been explored in the cognition and affect project as a move towards more human like systems.

A crude schematic representation of that option, the H-CogAff architecture (strictly another, more constrained, architecture schema), is shown on the right.

For more detail see the CogAff papers and presentations. Also Marvin Minsky *The Emotion Machine*

Supervenience etc.

Some philosophers have used an analogy between the mind-body relation and the value-fact relation. G.E. Moore, in his *Ethics* proposed that values supervene on facts: if two actions are factually the same then they must both be good or both bad – there's no ethical difference without a factual/descriptive difference.

This in itself does not imply that 'ought' can be derived from 'is', only that using words like 'ought' requires some consistency.

Donald Davidson: mental states or properties supervene on physical states or properties = two mental states cannot be different if the corresponding bodies are identical.

This in itself does not imply that the physical facts explain the mental facts, or that it is possible to derive mental descriptions from physical descriptions.

Much has been written about supervenience.

In particular there's no requirement that if two mental states are the same the physical states must be the same: most philosophers accept multiple realisability.

E.g. two people could both be thinking that the moon is smaller than the sun, but that does not mean that identical physical processes are going on in virtue of which they have those thoughts.

E.g. they could be using different languages, or different sets of neurones may have been recruited for that thinking process.

Problem: How can different physical processes implement the same mental processes?

What does this imply regarding the search for NCC?

Biological conjecture

I conjecture that biological evolution discovered those design problems long before we did and produced solutions using virtual machinery long before we did – in order to enable organisms to deal with rapidly changing and complex information structures (e.g. in visual perception, decision making, control of actions, self-monitoring etc.).

- You can't rapidly rewire millions of neurons when you look in a new direction
- or when you switch from approaching prey to deciding in which direction to try to escape from a new predator, using visible details of the terrain.
- So there's no alternative to using virtual machinery.
- But we know very little about biological virtual machinery.

  Nobody knows how brain mechanisms provide virtual machinery that supports proving geometric theorems, thinking about infinite sets of numbers, or algebra, or wanting to rule the world.

- The visual competences demonstrated here remain unexplained

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

We know that humans can use very different languages to say and do similar things (e.g. teach physics, discuss the weather); but evolution could not have produced special unique brain mechanisms for each language (since most are too new) – it's more likely that language learning creates specialised VMs running on more general physiological mechanisms.

Some conjectures about evolution of language are here:

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

Work to be done:
Biology, psychology, neuroscience, robotics

There is much work still to be done.

That includes finding out precisely what the problems were that evolution solved and how they are solved in organisms, and why future intelligent robots will need similar solutions.

There are more slide presentations on related topics here:

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

Many of the papers in the Birmingham CogAff project (Cognition and Affect) are relevant, especially papers on architectures.

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

But the problem of explaining how a genome can specify types of virtual machinery to be developed in individuals, including types that are partly determined by the environment at various stages of development is very difficult.

We need to understand much more about the evolution and development of virtual machinery.

See Jackie Chappell and Aaron Sloman, "Natural and artificial meta-configured altricial information-processing systems" `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609` (IJUC, 2007)

Closing Huxley's Explanatory Gap

If we can learn more about:

- varieties of virtual machinery and their roles in generating, controlling, modulating and extending behaviours in organisms;
- how they are implemented in various types of biological organism;
- how their features can be specified in a genome (e.g. the control mechanisms for mating, web-making, and eating in a spider seem, for many species to be genetically determined, although specific behaviours are adapted to the precise details of environment);
- how in some species the virtual machinery instead of being fully specified genetically is built up within an individual as a result of operating of genetic, environmental and cultural processes (see Chappell and Sloman, IJUC, 2007, mentioned above);
- how and why self-monitoring mechanisms came to include mechanisms able to focus on intermediate information-structures within sensory/perceptual sub-systems
  (e.g. how things look, how they feel, how they sound, etc.)

then we may be able to understand how a Darwinian evolutionary process that is already demonstrably able to explain much of the evolution of physical form might be extended to explain evolution of information processing capabilities, including the phenomena that lead to philosophical theories of consciousness.

But we should not expect there to be **one** thing, one **it** that evolved.

Darwin and his contemporaries knew nothing about virtual machines, alas.

Importance and implications of VMs

There are additional slides available on Virtual Machines, e.g.

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#mos09`
Virtual Machines and the Metaphysics of Science Expanded version of presentation at: Metaphysics of Science'09)

Topics include:

- Explanation of the importance of virtual machines in sophisticated control systems with self-monitoring and self-modulating capabilities.
- Why such machines need something like "access consciousness"/qualia – and why they too generate an explanatory gap – a gap bridged by a lot of sophisticated hardware and software engineering developed over a long time.
- In such machines, the explanations that we already have are much deeper than mere correlations: we know **how** the physical and virtual machinery are related, and what difference would be made by different designs.

Some remaining tasks

Perhaps explain how Freud's and other theories of multi-layered minds relate to all this.

Explain development of kind of self-awareness that enables people to tell the difference between what they treat as empirical generalisations and what they understand as (mathematically) provable – e.g. facts about topological relations, geometry, mechanics, and numbers. (The roots of mathematical thinking.)

A huge and important topic: disorders of consciousness, self-consciousness and control.

Much better understanding of nature-nurture issues, and requirements for educational systems.

John McCarthy on "The well-designed child".
`http://www-formal.stanford.edu/jmc/child.html`

Chappell and Sloman on "Natural and artificial meta-configured altricial information-processing systems"
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609`

Expand on varieties of metacognition, and differences between introspection and other aspects of metacognition.
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0803`

See other presentations in

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

and CogAff and CoSy papers:

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

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

Further Reading

Novelists have traditionally regarded themselves as the experts on consciousness, with some justification. See for example, David Lodge's essays and his novel on consciousness:
David Lodge, *Consciousness and the Novel: Connected Essays,* Secker & Warburg, London, 2002.
David Lodge, *Thinks ....*, Penguin Books, 2002.