Supervenience and Causation in Virtual Machinery

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

School of Computer Science, University of Birmingham, UK

This presentation is available at
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk86
and http://www.slideshare.net/asloman/presentations

Two closely related presentations:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk84
Talk 84: Using virtual machinery to bridge the "explanatory gap"
Or: Helping Darwin: How to Think About Evolution of Consciousness
Or: How could evolution (or anything else) get ghosts into machines?
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk85
Daniel Dennett on Virtual Machines

Related papers and slide presentations can be found at
http://www.cs.bham.ac.uk/research/projects/cogaff/
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/

Liable to change: Please save links not copies!

THIS IS WORK IN PROGRESS

Most of this presentation was written during 2010, though based on much earlier work.

The Meta-Morphogenesis (M-M) project, summarised in (?, ?) was begun late in 2011, as a result of my being invited to contribute to (?, ?), which led me to read (?, ?), which inspired me to ask "What would Turing have done had he lived longer?"

He was very interested in information processing in living things, including not only mental functioning in humans and other animals but also the control of growth of physical structures (e.g. the (?) paper).

I can't tell whether he anticipated the variety of types of virtual machinery that computer engineers developed during the remainder of the 20th century, but I suspect nothing in this presentation would have surprised him. He may have considered an even wider variety of types of virtual machinery.

There are two especially interesting comments in (?, ?) (a paper that is often referred to but rarely actually read, unfortunately).

> (1) Everything really moves continuously. But there are many kinds of machine, which can profitably be thought of as being discrete state machines.

> (2) In the nervous system chemical phenomena are at least as important as electrical.

A distinguishing feature of chemical information processing (e.g. in reproduction) is the mixture of continuous and discrete transitions (despite debates on discreteness of space/time in quantum theory).

I suspect, especially in the light of (2) that Turing was aware that this could have deep significance for understanding the evolution of biological information processing.

In the earliest stages of biological evolution there were no brains or neurons: the control systems were all chemical. Moreover even now, organisms with brains are mostly built under the control of chemical information systems, controlled largely, especially in the earliest stages, by DNA.

I suspect we shall find that chemistry can support important types of virtual machinery that are very hard, or even impossible, to implement adequately on digital computers, despite partial simulations.

This presentation may later need to be extended substantially to accommodate such advances.

The importance of virtual machines

Researchers in psychology, brain science, ethology, social science and philosophy refer to states, events, processes and entities in virtual machines when they talk about experiences, decisions, intentions, thoughts, learning, feelings, emotions...

The concepts used are not definable in the language of the physical sciences but they nevertheless refer to entities, states and processes implemented or realized in the physical world.

By having a clearer view of what virtual machines are, what they can do, and under what conditions they exist, scientists may come up with better, more complete, more powerful, testable, explanatory theories, using the "designer stance" (McCarthy).[*]

By clarifying the nature of virtual machines, their relationships to phenomena studied by the physical sciences, and especially their causal powers, we can shed light on old philosophical puzzles and explain why such puzzles arise naturally in intelligent, reflective, systems with human-like virtual machines – and why the puzzles will be rediscovered by future intelligent robots.

Some of these points are discussed in other presentations.
This one focuses mainly on issues regarding causation,
including "downward" causation from virtual to physical machine events.

[*]Dennett's "design stance" (mainly) attempts to understand existing systems.
McCarthy's "designer stance" requires building, testing, debugging, etc. and demonstrating feasibility.

Virtual machines are everywhere

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

– Poverty can cause crime
– Thoughts can cause desires
– desires can cause actions.

HOW?

Part of the answer is that they are all ultimately realized (implemented) in physical systems.

But these virtual machines are real and physics doesn't say what they are or how they work.

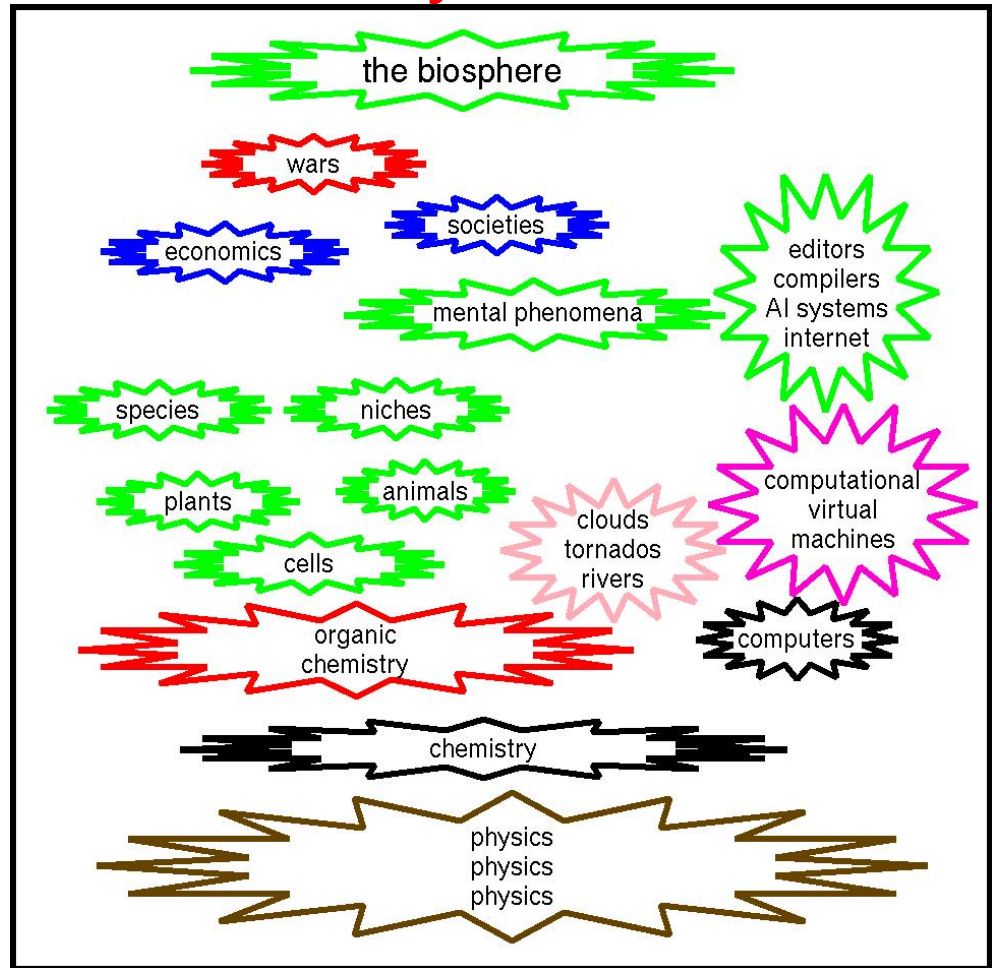Nobody knows how many levels of virtual machines physicists will eventually discover. (uncover?)

We'll discuss virtual machinery in brains and computers: both special cases of a general phenomenon: virtual machines can DO things.

What makes their causal powers possible is a combination of both physical properties of physical mechanisms AND very abstract patterns of interconnection instantiated in those mechanisms.

See also presentations on virtual machinery, consciousness and functionalism, in

    http://www.cs.bham.ac.uk/research/projects/cogaff/talks/
    http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html

Ocean waves as causes

Sea Storm in Pacifica, w:California 2008

By Mila Zinkova

Original image at

`http://en.wikipedia.org/wiki/File:`
`Waves_in_pacifica_1.jpg`

Used with thanks.

Two kinds of causation by a physical wave:

- An ocean wave can push you over, propel a surfer, move sediment on the sea bed, create high amplitude air waves.
  These are examples of **physical** causation by the wave.
- It can also remind you of a tsunami, make you want to go surfing, provide stunning visual and auditory experiences.
  These are examples of **using information received from the wave**.

Causation by information has been important for humans for millennia (many examples can be found in ? (?)), but our use of information in machines has accelerated enormously in the last half century, using new kinds of electronic machines.

Long before that, information played a multitude of roles in living things (?, ?).

I'll try to explain the differences between physical causation and causation in virtual machines that manipulate descriptive information and control information.

Three kinds of machine

"Machine" normally refers to a complex entity with parts that interact causally with one another and with things in the machine's environment.
   (That is not a sufficient condition for being a machine.)

Most familiar machines are Physical Machines (PMs)

Two other notions of machine, sometimes called "Virtual Machines" (VMs):
- an abstract mathematical object (e.g. the Prolog VM, the Java VM)
- a VM that is a running instance of such a mathematical object (RVM), possibly controlling events in a physical machine, e.g. a RUNNING Prolog or Java 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
PMs	**MMs**	**RVMs**

VMs that are mathematical models (MMs) are much studied in meta-mathematics and theoretical computer science. They are no more causally efficacious than numbers or axiom systems.

Many theorems in Computer Science, e.g. about computability, complexity, etc. are about mathematical entities (MMs), that are distinct from both Physical Machines (PMs) and Running Virtual Machines (RVMs).

Connections between MMs and RVMs

The situation can be confusing because theorems about mathematical structures (e.g. MMs)can be applicable to portions of reality which are instances of types of structure specified mathematically (PMs and RVMs).

- For example, ancient Egyptians used Pythagoras' theorem to help them solve practical problems relating to land management.

- Likewise a mathematical theorem about a MM can prove that an instance of the MM, a running virtual machine (RVM), will need at most a certain amount of memory.

- This is no different from using theorems about differential and integral calculus to prove things about motions of bodies in the solar system (a PM).

- In both cases, proofs can be relied on **only** if the non-mathematical objects actually have the properties used to define objects about which the theorems are proved.

- However, the world can surprise us: e.g. because planets do not move exactly as Newton supposed, or because a program has an unnoticed bug or because the hardware fails while the program is running.

- Much of theoretical computer science is about mathematical objects that are assumed to correspond to RVMs, but in fact may not always do so
  E.g. when there is a bug in the firmware of a CPU, as once famously occurred in an Intel design, or a bug in a compiler, or operating system.

RVMs instantiate MMs but are not themselves MMs: RVMs can make things happen or prevent things from happening. MMs cannot.

MMs have mathematical existence. RVMs have causal existence and depend on PMs.

Quine on what's real, and a better alternative

Quine: "To **be** is to be the value of a variable" (?, ?).

He equates what people think exist with what their quantifiers (e.g. "All", and "Some") range over.

This raises many problems, including the problem of how we can quantify over sets of entities about which we know nothing, e.g. future events, future people, or entities discovered in future by scientists.

Also: what about animals and young children who (presumably) don't use predicate calculus ?

A better answer: "Causes and effects exist" (More general than Berkeley's answer!)

We can make more progress if instead of talking about what exists, or is real, we talk about what is capable of being involved in causing things other than itself, or being an effect of other causes:

HAVING THE POTENTIAL TO INTERACT CAUSALLY IS SUFFICIENT FOR EXISTING.

That's one kind of being: numbers, sets, proofs, theorems, MMs. etc. have mathematical existence, which is concerned with being referenced in certain ways in mathematical and meta-mathematical theories and proofs.

There's also a related kind of existence in stories, plans, hypotheses, proposals, etc.

We may not know what caused a particular disease but if there is something that caused it then that something exists: **Being known to humans isn't a prerequisite for existence.**

We can hypothesize that a cause exists, even if we do not know whether it is a chemical compound, a living organism, a kind of hitherto undetected radiation, some psychosomatic process, etc.

Of course, if it turns out that what we thought was one disease is several diseases with distinct causes, then what we thought existed doesn't, but other things closely related to it do exist.

Objects, events, states, processes in RVMs are capable of being causes and being influenced by other causes: That is why we find so many virtual machines running on computers so useful — and why we depend on them more and more: they could not work if they did not exist!

(Even if most philosophers, including metaphysicians, have not noticed....despite using them every day.)

Another kind of existence: Mathematical

I wrote, above, that some things exist that are not causes or effects (though some of their instances can be): numbers, sets, proofs, theorems, MMs. shapes, structural relations, etc. have mathematical existence.

- To have mathematical existence is to stand in mathematical relationships to other things – to be consistent with, inconsistent with, to be a consequence of, to be possible in the context of, to provide a context for other things to be possible, or impossible.

- Some, but not all, of these mathematical relationships are purely logical: e.g. a set of well-formed logical formulae containing some undefined predicate-, relation- and function-symbols may form a deductive system with axioms, proofs from the axioms, and theorems that occur in proofs, where the proofs all use valid purely logical rules of inference. ("Purely logical" needs to be defined – but not here.)

- But long before that sort of mathematical existence was defined and studied (in the 19th Century onwards) there were also geometrical, topological, algebraic, and arithmetical structures, relationships, and forms of derivation that did not use the formalism of predicate logic

  as ? (?) implicitly claimed, in stating that mathematical truths are synthetic and necessary.

- We now know that mathematical existence can be relative to a mathematical system: e.g. some things exist in one system, but not another.

  Connected hollow objects with two or more holes are possible in 3-D spaces but not 2-D spaces.
  It's not clear whether there's a unique super-system of mathematical systems. (I suspect not!)

- Mathematical and causal sufficiency are deeply connected: a topic for another time.

Note on existence/reality

This note is a digression from the main point

- There have been many lengthy discussions and debates among philosophers regarding the nature of existence, or what counts as real.
- I regard many of the debates as futile because the questions are often posed as if they must have true or false answers whereas they are incapable of having such answers, for unobvious reasons.
  In contrast, questions like "How far away is the left edge of the universe?", "How many prime numbers between 20 and 90 dislike being prime?", "What was the last thing to happen before time began?" are obviously incapable of having true or false answers, apart from answers saying that the questions are incoherent, etc.
- Kant disposed of the ontological argument for existence of God (roughly "God must exist because non-existence is an imperfection and God by definition is perfect") by arguing that "exists" is not a predicate, and existence is not a property some things have and some things lack.
- Later work by logicians (especially Frege) showed more clearly the difference between predicates and the universal and existential quantifiers.
- Despite that, we can take "refers to something that exists (or is real)" as a predicate that can be applied to linguistic expressions (or their sense). E.g. the first one is true of the phrases "prime number between 20 and 30", "Oldest son of the current queen of England", and is false of "prime number between 24 and 28" "Seventeenth son of the current queen of England" (to the best of my knowledge).
- So one way to paraphrase the last few slides is to say that the expression "exists (or is real)" is correctly applicable to things that (roughly) either (a) are involved in causal connections with other things or (b) stand in mathematical or logical relationships within a certain deductive system without leading to contradictions in that system. In case (b) the existence/reality is relative to the system: compare Carnap on internal and external questions).
- We need to add informal deductive systems, e.g. stories, plans, hypotheses...
- There are still many questions not answered here, e.g. about the connections between the two kinds of existence, and whether the things that are involved in causal connections are all directly or indirectly related to one another (forming a single universe).

The realization/implementation/grounding relation

A FIRST DRAFT PARTIAL ANALYSIS: (? (?) makes related points.)

Phenomena of type X (e.g. biological phenomena) are fully grounded in, or realized in, or implemented in phenomena of type Y (e.g. physical phenomena) if and only if:

(a) phenomena of type X *cannot exist without* some entities and processes of type Y.

> (i.e. it is necessary that something of type Y exist for anything of type X to be a cause)

(b) certain entities and processes of type Y *are sufficient for* the phenomena of type X to exist – those entities constitute the implementation of phenomena of type X
NB. This is not logical or definitional sufficiency, though it may instantiate a mathematical necessity.
The **actual** implementation is usually not necessary for X: there can be alternative implementations.

NB: Such causal grounding has nothing to do with "symbol grounding" theory – criticized in
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk49

If computational virtual machines (RVMs) are fully grounded in physical machines then

(a) computational RVMs cannot exist without being embodied – implemented in physical matter;

(b) their physical embodiment (particular configurations of interacting matter) suffices for their existence: no extra independent stuff is needed – no computational spirit, soul, etc.;

(c) the particular physical embodiment P of RVM M is not necessary for M, insofar as M can in principle be implemented in different ways. (E.g. some parts could be replaced by new ones while M is running.)

(d) The language of physics cannot describe what is common to the sufficient conditions for M.

Explaining how P suffices for M often uses a collection of computer science concepts and a lot of know-how about hardware and software technology.

More on "fully grounded" (realized/implemented)

Phenomena of type X are fully grounded in, or realized (implemented) in, phenomena of type Y *if and only if:*

(a) Necessity condition: phenomena of type X *cannot exist without* the existence of entities and processes of type Y.

(b) Sufficiency condition: entities and processes of type Y *are sufficient for* the phenomena of type X to exist.

(Multiple realisability implies that quite a broad subset of phenomena of type Y will sometimes be sufficient to produce X – but there are constraints, mentioned below.)

The *kind* of sufficiency referred to in (b) needs clarification.

If a computer C runs an instance of a certain chess virtual machine VMc, we regard C as providing an **implementation** of VMc (and possibly other things at the same time).

- The configuration of physical processes in C is sufficient for VMc, though not necessary, as there are alternative implementations – including different possible implementations of VMc using C.

- Sufficiency for VMc is relative to certain conditions, e.g. the power supply is not turned off, the transistors in the CPU continue to operate properly, there is no destructive bombardment with gamma rays, no hitherto unknown physical phenomena disrupt the performance of the machine, no software intruder tampers with the operating system or some of the machine code, etc.

These ideas are close to philosophers' notions of supervenience e.g. (?, ?, ?), (?, ?), though they usually discuss supervenience of **states** or **properties**, not **virtual machines**.

The sufficiency may depend on conditions that cannot be specified exhaustively. Nevertheless it is a more than 'accidental' sufficiency, when the implementation works.

Some notes on sufficiency of implementation

- I stated that when a RVM M is implemented on a physical machine P, some physical parts of P required by M can be replaced by new parts (possibly even new parts with different physical materials) while M is running.

- This is not yet very common in computing systems but has become more common as a result of the requirement for "hot swappable" components of servers that must keep running even when components fail. In some cases, this depends on "redundant" implementations of memory and other components of computers, with quite costly duplication of function.

- This phenomenon can be compared with metabolism and related processes that repair and replace parts of plant or animal bodies without disrupting normal biological functions or mental processes, (though in some cases the immune system can disrupt normal functions).

  Maturana and Varela referred to this as "Autopoiesis".

- The biological cases are partly like the ongoing maintenance of a bridge while in use e.g. repainting and replacing weakened parts without interrupting the use of the bridge.

- This depends on a kind of redundancy: components work together in such a way that if a part is temporarily removed or disabled the rest can continue to provide their joint function.
  Compare "error correcting memory"
  `http://en.wikipedia.org/wiki/Error_detection_and_correction#Error-correcting_memory`

- We can expect metabolism-like processes to become more common in future man-made information processing machinery that needs to run non-stop.

  However, the possibility of doing that depends on use of very special modes of construction of all parts of the machines.

Sufficiency and conditionals

Saying that some state, event, or process (or type of state, event or process) is sufficient for something to exist has implications of two sorts.

"P is sufficient for V" has certain implications if P is false and other implications if P is true.

- If P is false then P is sufficient for V implies
  if P were true then V would exist.
- If P is true then P is sufficient for V implies
  of all the other things besides P that are true,
  if any of them were false and P were still true then V would still exist.

I am not saying that this is a definition of sufficiency.

That is because the concept "if ... then ..." used here can be thought of as itself expressing a notion of sufficiency, which would make the definition circular.

All I am saying is that the technical concept of sufficiency is a concept that is implicit in our understanding of the world, as exemplified in many (non-technical) uses of the "if then" construct.

Our ordinary ideas of causation are deeply connected with the truth of conditionals (counterfactual and categorical) – but that is not a definition of "causes", at least not for those who (like Kant) assume that where causal connections exist, something about the world exists that **makes** the conditionals true. Compare: ? (?), (?, ?)

So, we are not just referring to the truth of conditionals, but about what **makes** them true.

Example: Being in a closed room can make it true that if you start moving north you will be stopped. See also ? (?), for an introductory overview of some of the issues.

NB: I shall try to show elsewhere that "possible world semantics" cannot capture these ideas.

What sorts of things can be causes?

Not all philosophers believe that our ordinary notion of "cause" refers to something objective.

- David Hume famously claimed that he could find nothing in his experience to correspond to the concept of X causing Y other than the fact of X happening then Y happening, and that being an example of a true generalization, along with the existence in at least some cases of a very strong expectation that Y will follow X.

- Immanuel Kant argued that nothing can have experiences of an objective world unless it presupposes that there are causal connections between things in the world, and also between those things and perceptual experiences of them.

- Most people, including most scientists and engineers, ignore these issues and simply assume (implicitly) that causal connections exist and that in at least some cases they know how to recognize them, and how to make use of them, both in everyday actions and in engineering design.

- Among philosophers who think there are causal connections that exist independently of our discovering them, the majority view seems to be that only physical states, events and processes can be causes, even if effects may be non-physical, such as sensory experiences: this view I'll call "Rigid Physicalism".

- The same philosophers in their everyday life and most non-philosophers, assume that all sorts of non-physical things can be causes, including mental events, social events, economic events, and others. But it is not obvious how that could be possible.

Rigid Physicalism

What we could call 'rigid physicalism' states:

RP1: there is only one level of reality, the 'fundamental' physical level,

(I once heard a philosopher talk about "clunk-clunk" causation (e.g. as manifested in billiards), supposedly the only kind of real causation.)

RP2: causes can exist only at that level of physics, and nowhere else,

RP3: everything else is just a way of looking at those fundamental physical phenomena.

- Rigid physicalism leaves many questions unanswered: we have no idea what will be regarded as fundamental in physics in a hundred or a thousand years time, or perhaps is already so regarded by more advanced physicists on another planet.

- If the only 'real' causes are those that operate at the fundamental level of physical reality then our talk about one billiard ball *causing another to move* does not describe what is 'really' going on.

The extreme version of the theory that **'only fundamental physical causes are real'** implies that all our common sense beliefs about existence and causation are illusory.

Not only our belief that poverty can cause crime, that ignorance can cause poverty, that feeling sympathy can cause people to help others, that the spread of false information can cause social unrest, but also many everyday examples that we think of as physical causation such as ice on a road causing a crash would all be rejected as not being real causation.

The 'identity theory' of emergence: A defender of RP1to3 can claim that we talk and think about fundamental physical entities and processes even when we think we are referring to something else. I.e. economic inflation just IS some very complex physical process.

An alternative view: Flexible physicalism

An alternative view: causation is not restricted to a hypothetical 'bottom' level of physical reality, as there are different 'levels' of causal interaction.

Flexible physicalism allows that many things can interact causally: not only sub-atomic particles, force fields, chemical bonds, etc., but also more familiar entities, such as

> billiard balls, clock-springs, tidal waves, sunshine, epidemics, ecosystems, poverty, new ideas, being reminded, etc.

Including many non-physical things:

Economic inflation really can occur, and really can have effects. So can poverty, biological niches, social phenomena, economic phenomena, mental phenomena.

> A monkey and a parrot sitting on the same branch inhabit different biological niches.
> A niche can influence evolution.

Identity theorists agree that all these can be causes, because they all are no more than complex configurations of fundamental physical phenomena, even though we don't have any independent way of identifying the phenomena – a limitation of human knowledge.

I'll argue that there is a bigger gap than human ignorance, and will illustrate this by showing how networks propagating information can produce events that have causal powers that are not describable in the language of physics nor mathematically derivable from laws of physics. So both physical forces and information can be distinct causes.

Example: a Chess Virtual Machine

A computer user can have good evidence that a game of chess is being played (though the program's designer/maintainer has better evidence).

- The ontology of chess (kings, queens, bishops, pawns, rows, columns, diagonals, captures, threats, pins, etc.) is appropriate for describing what happens.
- However, the concepts in that ontology are not definable in terms of the concepts of the physical sciences. (That requires argument.)
- So it will not be possible to derive logical relations between the physical descriptions of what is happening in the computer and descriptions of chess playing.
- Therefore: From the fact that a machine satisfies a certain physical description, it will not be **logically** provable that the ontology of chess is instantiated in the machine – e.g. that a bishop is threatened.

Nevertheless, the chess ontology is instantiated because of what is going on in the electronic machinery.

- And that is not merely a subjective interpretation, based on personal preference.
- It is also not a mere contingent, empirical association: Designers know why those physical processes produce the chess-playing virtual machine processes.
- That's not just because they have observed a correlation.
- Moreover, some of them know how events in the virtual machine cause other events.
- Their work presupposes flexible physicalism.

Causal powers of ocean waves

On the 'flexible physicalist' view, many kinds of things other than fundamental entities of physics exist and enter into causal relationships, even if the existence and the causal powers of the other things 'ultimately' depend on underlying physical states and processes.

- Ocean waves can pound rocks, capsize boats, propel surfers, and disturb sediment.

- But the large and powerful ocean wave cannot exist or pound the rocks without vast numbers of tiny molecules moving around and interacting according to laws of physics:

  Wave motion is "implemented in" molecular motion and other things, including the earth's gravity.

- The same global features and effects could exist with very different molecular details:

  Different arrangements of millions of the water molecules might have produced the same global effects e.g. on sand patterns – but if the large scale patterns of motion in the wave had been different the effects on the sand would have been different.

- Large scale features produce effects that can be explained by Newton's laws.

- Unlike virtual machine processes (e.g. pawn capture), the large scale physical effects of ocean waves are aggregates of the sub-microscopic physical effects.

- The causation involves energy transfer from cause to effect, with most of the energy required to produce the effect coming from the cause.

Something different happens when information acts as a cause

(as ? (?) noticed).

E.g. when an intrusion-detection system notices a face that it does not recognize and sounds an alarm.

Causal powers of **patterns** in ocean waves and in sand

Contrast waves pounding a shore with:

Someone watching a wave, and being reminded of a tragic tsunami.

Perceiving the pattern of motion of the wave causes memories to be revived.
Likewise seeing a static pattern in the sand can bring a forgotten face to mind.

There are similarities and differences between pounding and reminding:

- A pattern in wave motion and a pattern in static sand cannot exist without physical material,
- but the pattern is not just the material.
- If the sand had been rearranged and the pattern removed it would not have had the same reminding power: the pattern has effects that the material alone lacks – e.g. triggering memory processes.
- The sand could be replaced by many substances with very different physical properties, e.g. sugar, salt, lead pellets or mud, in the same pattern, and the pattern could still trigger the same memories.
- Many details of the wave motion could also be changed without affecting the ability of the perceived pattern to remind someone of the previous tragic event: the reminding pattern can be very abstract.
- **Much less energy needs to be transferred from the cause to the perceiver for reminding to occur: most of the energy required comes from inside the perceiver** –
it is information not energy that has the main causal power to remind, in this case (?, ?).

The moving water causes rearrangements of sediment by acting on it and transferring energy, whereas memory processes in an observer are triggered much less directly.

For reminding effects of a wave pattern, or the sand pattern, to occur, very little energy needs to be transmitted: what's important is not the amount of electromagnetic energy but the spatio-temporal pattern in the light: it is **structure** that informs.

NOTE: The reminding pattern can work even if filtered through coloured glass, or received monocularly.

Being reminded vs. being pushed

- The sediment, rocks, and marine organisms pushed around by waves are directly affected by the forces: only very general laws of physics are involved in the process.
- When seeing a wave or a sand pattern reminds you of something, electromagnetic energy received triggers vast numbers of processes in your brain that detect myriad information fragments (about edges, textures, colour, various kinds of motion ....).
- Those information fragments (pattern parts) are actively grouped in various ways, and larger wholes assembled – often with alternative groupings competing.
- Some of the patterns of activity trigger reactivation of previously acquired information structures, some of which then influence subsequent grouping and inference making.
- In a very short space of time various structures, processes, and relationships in the environment are represented and this causes new patterns of information to flow to various parts of the brain: using up considerable amounts of chemical energy.
- A common side-effect is to re-activate previously stored items of information (about a particular tsunami, or person, or unanswered question, or story read last week, or ...).

  The perceiver and the received energy both contribute to the result in ways that depend hugely on both the information-processing powers and the previously stored information in the perceiver: a young child or a chimp will not see what you see, and another adult human will not be reminded of the same things. Even the same person will be affected differently at different times.
  Likewise, an intrusion-detection system can learn about new "good" and "bad" faces.

- General Laws of physics and chemistry are not adequate to explain what goes on.
- But it's not an inexplicable mystery: there is a (growing) science of information.
  Including a (still very primitive) science of visual information processing ? (?, ?).

How can information have causal powers?

What we have learnt in the last half century about putting information to work, riding on the back of physical causation, is very complex and not easy to explain.

But one of the most basic phenomena has been known about and used for a very long time: the ability of a pattern to cause **itself** rather than **another pattern** to be transmitted, using replicating machinery. (? (?) gives many historical examples.)

Consider the production and use of a piano-roll recording of a piano performance.

- The actions of the pianist instantiate an abstract pattern in a temporal sequence of chords (combinations of pitches – since notes can be played concurrently).
- Physical mechanisms sense the abstract pattern in the motions of the piano keys and, possibly using an additional power source, transfer the pattern to holes in paper.
    (Using a static 2-D spatial pattern to encode aspects of a 2-D chordal temporal pattern.)
- Copies of the paper pattern can be made and later one of them fed into a "player piano" mechanism which uses the sensed spatial pattern to generate a pattern of piano hammer motions, producing a new instance of the original acoustic pattern.
- In principle that automated performance could be recorded and used to produce yet another instance later.

The musical pattern is an **abstraction** that can be instantiated in different physical media at different times and can cause itself to be replicated by being fed into one of several types of machinery that take in and produce patterns.
(Bit-patterns are a very simple, special case. ? (?))

Information, unlike energy, is re-usable indefinitely

The energy in a wave, or barrel of oil, or battery, can be used to produce effects, but in the process it is also used up (more strictly converted to an unusable form), so use of an energy store diminishes future use.

But Information items can be re-used indefinitely without being diminished.

Three main causal roles for information:

1. Arrival, or internal derivation, or generation, of new information can trigger new events, including new inferences, new goals, new actions, new storage processes, across various time scales.

2. Change of information can trigger new events (Also over various time scales.)
   Detecting change requires that something new has occurred, but also requires mechanisms that can compare old and new items.
   (So it is not **change blindness** that needs to be explained but **change detection** – and its failures.)

3. Previously acquired static information (e.g. linguistic knowledge) can control or support new inferences, or constrain effects of new information or effects of new goals or preferences: Static information can go on having such effects indefinitely,

In computers, many mechanisms, some hard-wired, some in "firmware", some in software, some in attached devices, have been designed to ensure that such causal connections exist, some permanently, some turned on and off by other mechanisms.

Brains too, it seems, though we don't much about details of virtual machinery involved.

Note:
I am not using "information" in the (syntactic) sense introduced by Shannon in discussing communication mechanisms (?, ?), but the much older sense e.g. used by Jane Austen in her novels, illustrated in
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/austen-info.html`

Pattern Propagation

The most primitive form of causation by information is pattern copying, which can be used for propagation along a pipeline composed of a chain of pattern readers, writers and storers.

- The various instances of a pattern in a pipeline could have different physical forms, using different media:
  - some dynamic (an acoustic pattern, a pattern of movements of piano hammers, or more recently patterns of electrical pulses, or patterns of photons)
  - others static (e.g. a pattern of holes in paper, or more recently a pattern of arrangements of magnetic molecules on a tape or disc, or a pattern of microscopic deformations on a reflective plastic disc).
- In a particular propagation pipeline, what happens is caused by a combination of:
  - (a) The fixed physical configuration of the mechanism capable of propagating different patterns, including its energy supply.
  - (b) The actual pattern that is fed into the system – an abstraction that can be instantiated in various different static and dynamic physical forms, at different locations in the pipeline.
- In order to be physically storable and copyable the pattern must be capable of having physical instantiations with properties that are sensed physically and cause specific features of the reproductive process.
- The pattern can include quite different abstract properties that are preserved across the different physical manifestations,
  - like a set of equally spaced rows of dots also having columns and diagonals,
  - or the polyphonic structure of a fugue in which several melodies (voices) are concurrent – properties to which the physical copying and storing mechanisms are insensitive.

From Propagation to Interaction

- When a pattern causes itself to be replicated or propagated, it can be regarded as having purely self-referential semantics: it denotes itself and its instances.

- Patterns can not only cause themselves to be propagated, but can also have many other effects, especially when they interact with other processes.

    Creating a sequence of stacked rows of dots can also, as a mathematical by-product, produce vertical columns of dots and also diagonals made of dots.
   Replicating the row of dots also necessarily replicates the columns and diagonals.
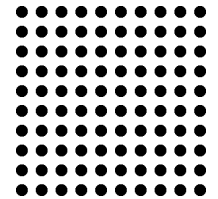
- If the dots are printed in different shapes, e.g. a circle, a triangle, a square, etc., then that feature of the replication process can interact with what is copied (a fixed number of dots in a row) to produce a column of circles, a column of triangles, etc.

- If another pattern, an oscillatory signal is sent to the row of printing heads in parallel with the instructions to produce the row of dots, then by moving from side to side the printing mechanism will produce wavy columns: a new 2-D pattern, of waves, arises from the repeated replication of two other 1-D patterns. (As in a seismograph.)

- The patterns in a piano-roll directly cause a single temporal sequence of physical events to be produced, but, like the columns and diagonals, the replication process may produce, as a by-product, polyphonic music, e.g. with two parallel, interweaving, tunes, or even a multi-voice fugue.

- Similarly the patterns of holes in punched cards driving a loom can, in combination with coloured threads, produce complex 2-D patterns in cloth that are inevitable consequences of the punched-hole patterns, given that loom's physical mechanisms.

From Copying to Detecting and Remembering

- We have seen that physical mechanisms can have patterns fed in which they replicate, while another pattern fed in modulates the replication process.
- There could be not just two, but several, streams of incoming patterns interacting in complex ways, producing multiple output streams.
- If everything is held constant in some way, apart from one of the input streams, then the behaviour of the whole may be caused to change in a characteristic way when that input stream changes: i.e. the change is detected.
- Some streams may have their contents modified by other streams merging with them, or can be blocked and turned on by other streams.
- If there are multiple branching pipelines, each modulated by inputs from other pipelines, where some input streams come from external sensors, and some patterns are transmitted to controllers for external motors, then a fixed information-processing architecture has changing behaviours, influenced by current and recent inputs.
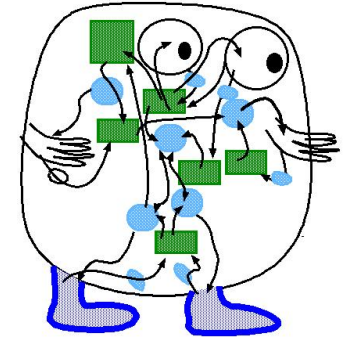
    Several such systems, where all the sensory patterns and motor signals have scalar (numerical) values, a very simple special subset, are presented in ? (?).

- Some externally induced patterns can stimulate circular pipelines that keep the pattern cycling or reverberating, so that their influence on other subsystems endures.

    So information about a previously detected feature of the environment can go on influencing internal and external behaviour – possibly using energy while retaining the information ready for use.

    Further complications of varying kinds would correspond to increasingly sophisticated animal-like features, including diminishing energy stores that need to be monitored and occasionally replenished.

... and resisting, or competing

- A piano is built physically so that if left alone it returns to a state in which no keys are depressed and no hammers are hitting strings.
- Attempting to play the piano involves depressing keys, which are built to resist depression – but not too much: the performer usually wins.
- In more complex mechanisms exactly how the influence of one pattern is resisted depends on the influence of other patterns.
- It is possible for two patterns to compete over a physical state (e.g. keys and hammers at rest or moving), while competing over which new pattern is produced: one person attempts to play a piano while another repeatedly tries to play something different on the same piano, including resisting the movements produced by the first player.
- Compare two chess programs running on the same computer and competing

    each attempts to make moves that will lead to a win, and each attempts to block the other's strong moves and make the other player lose: neither can directly modify the other's internal processing, but changing the (virtual) board changes what the opponent can do or needs to do.

- The ease or difficulty with which each program can do what it is trying to do can change according to the state of the game.
- Resistance of one pattern/program to another and the influence of one over another can both depend on the details of the physical implementation, sometimes in very complex and constantly changing ways.

    E.g. in some cases the competing subsystems are **concurrent**, unlike the chess example.

- Biological information processing often involves competing virtual machines and also virtual machinery that detects and resolves some of the conflicts.

Competing with varying strengths Added: 4th Dec 2010

- When two or more physical objects try to push the same thing in different directions, it will move as if pushed by a single force: the resultant, i.e. the vector sum, of the forces.

- If two directly opposing forces of different magnitudes are applied to **X** then **X** will move in the direction of the stronger force, its acceleration depending on the difference between the two forces.

- But competing causes in an information-processing VM are not like competing physical forces:

  e.g. a chess machine may be composed of two competing sub-systems, one cautious and one adventurous and which one determines the next move may depend on the state of the board and perhaps the system's estimate of the competence of the opponent: but only legal moves can be selected, never something like a vector sum of legal moves, which might cause a piece to straddle two squares. (A young child playing with chess pieces can take some time to develop such constraints.)

- Much computing/AI research since the 1950s was/is concerned with systems that select among alternatives, using a variety of different mechanisms, including
  – Searching for an option that satisfies some criterion (e.g. achieving a goal);
  – Searching for an option that is optimal on some measure (e.g. closeness to a target);
  – Learning scalar measures of competing sub-systems, and selecting the one with the highest measure to produce the option (consult the best expert);
  – Using a "relaxation" method – iteratively allowing competing subsystems to make small changes until a static equilibrium state is reached;
  – Using a random mechanism (e.g. as a last resort – like a human spinning a coin or throwing dice);
  – Being "fair" and repeatedly giving all options a turn at winning (like a scheduler);
  – Using a rule-based decision maker to select what to do, taking into account features of the options and the context, and perhaps results of previous selections. (The methods need not be numerical.)

A different notion of strength (5 Dec 2010)

It should be clear from the previous slide that when there is competition between alternatives in an information processing system, the competition can be very different from processes involving opposing physical forces.

An important research problem is to understand the different forms of VM competition and different types of conflict resolution mechanism that are possible and useful in biological organisms – and future robots.

There is still much work to be done exploring the role of self-monitoring and self-modulating systems and ways in which detection of conflict can usefully alter subsequent processing. (Compare ? (?, ?))

People who deny that computers can have experiences like human or animal experiences may base this on an intuition that things like pleasure, pain, being torn between two options, desperately wanting something, regretting a past action, are states that cannot possibly occur in computers. They forget that these are control states (?, ?).

Insofar as the issues being discussed here have not (yet) been addressed adequately in AI, those intuitions are not mere prejudices against AI (even if they are wrong).

It seems that some forms of internal conflict and experience of conflict require truly concurrent systems that can monitor and try to alter one another in parallel.

Battles on the internet between software systems attacking and defending web sites may be better models than the current shallow models of emotions in AI/Robotics.

The internet battles are real unlike the battles that a fake emotional robot may report in words "I am torn between..." (Compare (?, ?)).

Models of mental conflict (5 Dec 2010)

Over past centuries there have been many ways of thinking about how minds work, including how internal conflicts arise and are resolved, e.g.:

- Using the analogy of competing physical forces –
  But unlike actions of forces the outcome can take time and is not usually a vector sum or "resultant".

- Regarding the mind as something like a parliament or society of competing factions.
  E.g. Plato and Minsky and Bible ("It is sin that dwelleth in me".)
  This model risks circularity if the combatants are thought of as human-like.

- Inventing special parts of the mind that have their own forms of behaviour and interaction to be studied by introspection, or empirical psychology.
  E.g. Freud's id, ego and superego, various notions of self and its internal opponents (e.g. instincts, habits), "the will", reason, animal tendencies, drives, etc., multiple-personality theories, and more

- Supernatural theories about souls, demons in possession, gods that intervene, etc.

- Materialist theories referring only to brain mechanisms and processes,
  e.g. hormones and other chemicals, neurones and neural interactions, non-local quantum interactions, mechanisms to be discovered in future, possible telepathic mechanisms (like radio signals?) etc.

- There are differences in the various theories according to the what the conflicting entities are. (?, ?, ?) Conflicts occur between
  – What to attend to (selection between alternative perceptual contents, or alternative previously adopted tasks).
  – Alternative goals to aim at.
  – Alternative physical behaviours.
  – Alternative ways of processing some sensory or other information.

We still need good models of such conflicts and ways of resolving them. (?, ?)

Networks of networks of networks... of influence

Familiar neural nets are special cases of information propagation devices whose operations are controlled or constrained by information

e.g. encoded in network topology or in connection strengths or polarities, or in levels of activation.

There are many special cases:

- Some networks are continually changing – cycling through various states (with or without repetition).
- Some are static much of the time, but multi-stable, so that small influences can produce rapid changes to new static states.

  Computers use very large numbers of bi-stable elements, linked in carefully designed networks.

- Some networks are composed of smaller networks, not necessarily all similar in structure and function.
- A network can be composed of networks, composed of networks, composed of ...
- Some of the changes triggered can alter the topology of a network, adding or removing some subnets or connections between subnets.
- Some of the sub-nets may be closely connected to sensors and/or effectors (motors) so that they continually interact with the environment, while others go through changes that depend more on the history of the total system, or even the evolution of a species, than on current interactions with the environment.

  Disconnected sub-nets are more suitable for imagining, remembering, planning, forming goals, deciding preferences, constructing hypotheses, reasoning from available information, ...
  Temporary connections can create temporary sensorimotor influences.

- Patterns processed may be scalar values, or relational structures, e.g. molecules.

Fixed vs Changing Information-Processing Architectures

- It is sometimes possible to identify fixed subsystems linked by communication channels in such a network: the fixed topology and set of functions and connections constitutes the architecture of the system.

- An architecture can be described at a level of abstraction that allows different instances to differ in the details of what they do, how they do it, and which physical mechanisms they use.

- Many information-processing systems, especially human-engineered systems, have a fixed architecture, though not all: for instance the architecture of the internet has been constantly changing, in several different ways, over the last two decades.

- Biological information processing systems cannot have a fixed architecture: there is no designer, with a factory, to produce them: instead they grow themselves partly controlled by a specification in genetic material.

    In some species (e.g. precocial species) that always produces roughly the same adult architecture.

- More sophisticated organisms (altricial species) grow their architectures in ways that can be heavily influenced by the environment (including teachers), possibly adding several different layers of structure, each depending on what is achieved by previous layers interacting with the environment. (?, ?)

    Nobody knows how human architectures develop and what all the genetic and environmental influences are. But clearly there are many stages, adults are both very different from themselves as infants and from other adults, especially adults in very different cultures – e.g. stone-age humans.

- It is also possible to change sub-architectures rapidly, to suit different contexts.

Conditional Modification and Propagation

A network full of relays for transmitting patterns,
- where individual transmitters can be temporarily turned on or off
- or speeded up or slowed down,
- or caused to modify the patterns transmitted,
- on the basis of modulating patterns received by those transmitters

may be capable of enormously varied patterns of internal and external behaviour,
partly influenced by the environment, on varying time-scales.

A special case is John Conway's "game of life", using a 2-D grid of bi-directional pipelines
transmitting single ON/OFF patterns, each node obeying the same set of simple rules.

Play with Edwin Martin's online implementation here: `http://www.bitstorm.org/gameoflife/`
or this (larger) one `http://www.conwaylife.com/`

Learn more about "Life" here: `http://en.wikipedia.org/wiki/Conway's_Game_of_Life`
or here `http://www.conwaylife.com/wiki/`

- Conway's "Life" is restricted to (a) deterministic rules, (b) binary states, (c) 2-D, 8-way
  connectivity, (d) each cell's state determined by immediately preceding states of
  neighbours, (e) synchronous operation, and (f) no external inputs while running.
  (Though some implementations allow mouse-clicks to alter cells in a running system).

- These specifications and descriptions of particular systems are
  implementation-neutral.

- By relaxing restrictions (a) to (f) we obtain a very much larger space of possible
  deterministic and non-deterministic information-processing networks, not all of which
  can be modelled on a Universal Turing machine.

Generalising dynamical systems theory

- Systems described here generalize the common notion of a dynamical system whose state can be represented by a (possibly very large) vector of scalar values (e.g. real numbers) and whose behaviour is wholly describable in terms or the ways the numbers change,

  > e.g. using differential equations for continuous change or difference equations for discrete dynamical systems whose components all change in step.

- The generalization above allows complex dynamical systems, composed of changing collections of sub-systems, whose states and processes are not restricted to scalar values and quantitative changes: for they can contain static or process patterns of many sorts, some of them more like shapes (with parts and relations), or graphs or trees, or looping processes – not just numbers or bit patterns.

- Partitioning of a system allows us to refer not only to state-changes of the whole system, but also to influences between of sub-systems and states, state-changes and trajectories of sub-systems, temporary and enduring.

- Environments are also connected sub-systems, or sets of sub-systems (?, ?)

- One subsystem, or collaborating group of subsystems, may transmit information to another, or modulate transmissions between others, or monitor other subsystems and transmission channels, or cause new groups to form (Compare (?, ?).)

Physical and Non-Physical Describability

- In some cases the interactions can be described as a sub-system selecting a goal state, and recruiting other sub-systems to drive the whole into the selected goal state.
- Subsystem A can be trying to achieve state S for another subsystem B insofar as A monitors states of B and when B moves out of S, A takes action to steer it back into S, and also takes action to prevent other things that attempt to change B from state S.
- The preserved state S may be defined in terms of abstract relationships between patterns of activity in the whole system, where it is immaterial what physical processes implement the state.

    The state being preserved is defined (for the system attempting to preserve it) only in terms of relationships between abstract sub-states.

- The actions taken by A can be more or less direct: it may be able to directly manipulate some of the components of B, or may be able only to invoke other subsystems with instructions to alter B into the required state S.
- There need not be any determinate physical description of the state that is being monitored and preserved since the existence and influence of that state depend only on very abstract relations of causal influence –

    just as the causal relations between states of chess playing program are describable only in terms of concepts of the game of chess, which are independent of any particular physical medium, even though some physical medium must be involved whenever the game is played.

- This illustrates how an abstract description can be true of a complex dynamical system because the network of causal interactions within the system supports a changing network of conditional and counter-factual conditional truths about the system.

Physically-describable and non-physically-describable

Every network supporting acquisition, storage, propagation, transformation, and use of information, based on the mechanisms described above, uses a physical machine whose detailed operations are fully describable in the language of physics (including mathematics).

But some networks may generate patterns of interaction that are best described using a different ontology, such as the ontology of chess, where patterns satisfying that description could exist in a variety of different physical configurations of machinery.

Compare the different collections of physical fragments, of sand, salt, sugar, ..., that could instantiate the same face-like pattern.

Some patterns have their effects because of their relationships to other patterns, which have relationships to yet more patterns.

A pattern may be implemented in different ways, as long as the causal relationships are preserved: the same conditional and counterfactual conditional statements remain true.

The ontology of physics does not include concepts like "goal", "trying", "attacking", "protecting", "rules", "legal move", "following rules".

There need not be definitions in physical terms of what those patterns are or what they do.

In that case there will be no way of proving logically, from a purely physical specification of a machine that it implements the virtual machine in question – even if it does.

Note on indefinability

It could be argued that the allegedly physically indefinable abstract causal networks (RVMs) are physically describable as follows:

Form a disjunction of descriptions of all the physically possible instantiations of the abstraction that would constitute an implementation of the RVM.

That could be a **definition** of the type of RVM implemented.

There are many objections to this: e.g. how can we tell when we have a complete specification of the set of possible instantiations, including all those that could make use of future technology?

What concept of the RVM are we using when we seek and then discover or create a new physical implementation? Do we then abandon our old concept and replace it with a new disjunction?

How can we justify the inclusion of a new design into the disjunction? Surely that requires us to have had a prior understanding of what the new design is a design for?

The old concept cannot be defined by the collection of previously known implementations, since the new implementation would not fit that concept.

(to be revised... Can Ramsey sentences provide a translation?)

I'll now show one of the ways in which we can assemble evidence regarding the nature of virtual machinery required to explain some human competences, especially animal visual competences.

Multiple layers of information in perception

It is obvious and widely understood that understanding spoken or written language, or a musical performance, requires information at different levels of abstraction to be detected, constructed, inferred and used concurrently: similar complexities occur in visual and other forms of perception.

They are all cases of "multi-window perception". (? (?, Ch. 9),(?, ?))

- A complex and changing pattern in incoming energy received by sensors (acoustic, optical, haptic, etc.) can, as a result of previous processes of biological evolution, learning and development, activate a host of information-processing mechanisms that analyse and interpret the signals in terms of different kinds of information content, at different scales, constructed in parallel, in cooperative processes.

- This contrasts with "peephole perception" where a sensor stream goes through a sequence of mechanisms that analyse or interpret the outputs of previous stages, some of which produce other changes as side-effects (in memory, in current goals, in motor signals). (?, ?)

- It also contrasts with the simple-minded dynamical systems view, often linked to an emphasis on embodied cognition, assuming that **all** information processing is closely coupled with sensory-motor signals crossing the organism-environment boundary, as crudely depicted on the right:

- In a "multi-window" sensorimotor system there can be many different dynamical systems, some dormant, others operating concurrently but asynchronously, on different time scales,

- using information concerned with various aspects of the environment, not all of them currently sensed or acted on. (See the next slide for a contrasting sketch.)
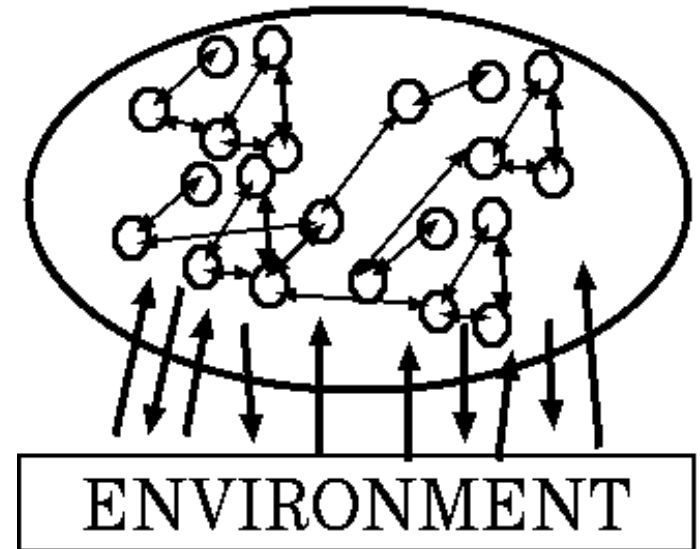
ENVIRONMENT

Multi-level dynamical systems

An alternative view of the CogAff Schema

(Described in
http://www.cs.bham.ac.uk/research/projects/cogaff/)

In these "multi-layer" dynamical systems, only states and processes in sensorimotor sub-systems (bottom of diagram) are closely coupled with the environment through sensors and effectors, so that all changes in those layers are closely related to physical changes at the interface.

Other subsystems, operating on different time-scales, with their own (often discrete) dynamics, refer to more remote parts of the environment, as indicated crudely by the red arrows: e.g. referring to internals of perceived objects, to past and future events and places, to objects and processes existing beyond the current sensors, or possibly undetectable using sensors alone (using an exosomatic ontology.

Those decoupled subsystems can refer not only to what is happening but also to what might have happened, could be happening out of sight, or could happen in future, and also to constraints on possibilities. (Proto-affordances and various kinds of affordance.)

Besides exosomatic factual information (referring to entities beyond "the skin"), some of the higher level subsystems can include **questions, motives, preferences, policies, plans, and other control information, also referring (amodally) to more or less remote, or past or future entities.**

Some new items of information will produce only minor perturbations of on-going processes (e.g. tracking motion), others major disruptions (e.g. detecting a serious new threat or obstacle).
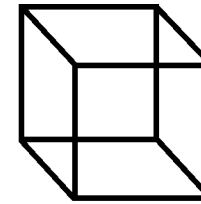
ENVIRONMENT
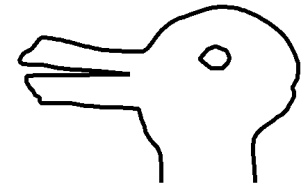
Remote and hidden parts

**Crude sketch of
some aspects of seeing**

Machines for multi-level minds

Analysing what flips in ambiguous figures suggests
some requirements for mechanisms of mind.

Necker Cube   **Duck-rabbit**

"Rows? Columns?"

"Diagonals?"

- Some aspects of seeing seem to require analog processing,
  with continuously changing quantities and control loops,
  e,g, seeing continuous rotation. pulsation, hand-shaking, ...

- Other aspects suit digital information processing, using
  discrete operations, e.g. distinguishing "F" and "E", or rows and columns.

- For some parts of the multi-layer system we need sub-systems that are multi-stable,
  and can be either active or dormant: when active, information from other sub-systems,
  or possibly even noise, can tip them from one stable state to another.

- If there are large numbers of such systems linked together, spreading patterns of
  activation can trigger rapid reorganization, including re-grouping.

    Ripples caused by small stones dropped into a pond illustrate one kind of multi- component dynamical
    system, but ripples and waves can pass through one another: we need much richer interactions.

- Instead we need something analogous to hordes of map-makers, meccano-model
  builders and circuit designers, who, if prompted, can rapidly reassemble the
  components they handle, and allow them to join up with other components, e.g.

    – Dot descriptors can be assembled into row, column, diagonal, square, diamond ... descriptors.
    – 3-D fragments can be assembled in different orientations, and different relative distances, for
      different views of a wire-frame cube.
    – Fragments of other kinds are joined up to form competing internal percepts of a duck and of a rabbit,
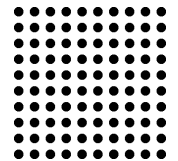   including different directions they face towards.

Don't assume representations must be replicas

- If I give you a banana you can eat it – and then you have it no more. If I give you information about where a banana is, you can't eat the information, but you can do other things, including giving the information to others while still retaining it: something you can't do with the banana.

- Requirements for, and uses of, information about X usually differ from requirements for, and uses of, X itself: A cave, but not information about it, can shelter you from rain.

  Don't be misled by overgeneralizations like "the world is it's own best representation", ignoring the fact that information about a cake (e.g. how to make it) serves purposes that the cake cannot and vice versa. You can carry information about the location of a building, but not the location itself.

- What is true is that in some special situations we can use servo-control (feedback from the environment) instead of ballistic control based on pre-stored information.

  E.g. looking at or feeling whether your fingers are in the right location to pick something up, instead of trying to work out in advance precisely how to move, using only previously acquired information.
  – It is often useful to allow a portion or aspect of the world to play a significant role in controlling actions relating to that portion of the world, but those are special cases.
  – If you want to build a large dam you need a great deal of information about where the dam is to be, what shape it will have, etc. Since the dam does not yet exist it cannot give you that information.
  – Even if you know another dam exactly like the one you want to construct, simply looking at it will not necessarily give you much information about the processes required to construct another.

- We need to analyse tasks of different kinds, to learn what sorts of information can be useful and how they can be useful for those tasks, possibly after reorganisation.

  Fragments reorganised may be much more abstract than maps, meccano pieces, or electronic circuits.

Assembling fragments of spatial information in a spatial way

Spatial information in non-spatial structures

It is often assumed that when a 3-D scene is perceived the brain somehow constructs either an isomorphic 3-D replica or something that is mathematically equivalent to such a replica, and can, for example, be used to project 2-D views of the scene in various directions.

(Often used as a test for success of 3-D stereo vision in machines.)

This ignores the fact that what an animal (or robot) needs in order to control behaviours, or to reason about possible actions and their consequences, is usable information about what is in the scene; and what is usable will depend on what it will be used for.

Reutersvard Triangle (1934)

Deriving possible consequences of picking X up, or throwing X into a river, is different from actually picking X up, or throwing X into a river. (Off-line vs On-line intelligence.)

Working out the consequences need not change X's location, or cause something balanced on X to alter it's location, or cause X to be irretrievably lost in the river.

Information structures built in perception need not be coherent – though normally they will be, because based on information from things that exist – unlike the scene above.

Information can be assembled in a manner that reflects some of the spatial structure of what the information is about even though the information does not have exactly that spatial structure. **X is represented as further than Y** is different from **X is further than Y.**

Control information and other kinds

The most primitive form of information is control information: specifying what to do by triggering activation of one piece of (internal or external) behaviour rather than other available alternatives:

Pressing a button on a vending machine (after inserting money) can trigger control of the vending process: which item should come out.

but that's not the only form of information.

- Storing information about the prices of items in the machine does not directly generate behaviours, but it can help to control choices between various sorts of internal or external behaviour.

- The same information item can cause different behaviours in different contexts

- specifying the price of a cup of coffee as 60p can have a variety of contextual effects:
  - comparing the amount of money inserted with the price, in the context of coffee being selected.
  - requesting more money if the amount inserted was 50p
  - causing the coffee to be poured and 40p change to be given, if the user has inserted 100p.
  - .... etc. ....

An important feature of such processes is that when a particular new item of information arrives, what internal and external processes occur can depend on the state of the receiver, in particular what other items of information have previously been acquired.

How that state is implemented may not matter, as long as the truth of a web of conditional statements ("IF p&q&r THEN s&t" etc.) is preserved by well designed chains of influence.

Physical and informational causation

A major difference between familiar physical causation and informational causation is concerned with the source of energy used.

- When rotation of one gear wheel causes another to rotate, or when a marble is caused by gravity and the shape of a helter-skelter to roll down in a roughly spiral trajectory, the caused behaviour uses energy from the causes:
  - One gear wheel transmits energy to another
  - The earth's gravitational field and surfaces of the helter-skelter apply forces to the rolling marble. (Some forces are elastic responses to impacts from the marble, whose gravitational potential energy is transformed to kinetic energy temporarily stored as elastic energy by surfaces it bounces off.)

- When organisms or machines use information to select options for action or for further information-processing, the information transmission may use miniscule amounts of energy compared with the energy deployed in the (internal and external) actions.
  - Small amounts of electromagnetic energy reflected by visible surfaces often allow hugely complex collections of items of information about the environment to go via retinas and brains into minds.
    A lot more energy is used by the brain in interpreting that information.
    Even more energy may be expended following a decision to run away from approaching danger.
  - Likewise information received by a robot via its cameras or other sensors involves transfer of small amounts of energy, but the information can lead to decisions that initiate actions using far more energy available in the robot's batteries or other energy supplies (e.g. shutting and locking doors).

- In computers and brains there are many information-processing steps, combining information from many sources, with varying delays between receiving information and using it to select actions: unlike propagation of physical forces through a machine.

It's an old idea

The idea of using a small amount of energy to control the deployment of a large amount of energy is very old, in mechanical devices.

- A Watt governor siphons off a small amount of the kinetic energy produced by a steam engine to measure the speed of the engine, by rotating a device whose location alters with speed, and can thereby control a throttle on the steam input pipe.
  Information about the actual speed is used to control speed in the immediate future.
  `http://en.wikipedia.org/wiki/Centrifugal_governor`

- A low power device can operate a ratchet mechanism that controls the speed of an axle transferring much more power. `http://en.wikipedia.org/wiki/Escapement`

- The vane on a windmill uses a (relatively) small amount of wind energy to obtain information about direction the wind comes from, and rotates the main blades to capture far more energy to drive the mill.

- In a Jacquard loom, small amounts of energy used to detect positions of holes in punched cards can control the use of far more energy driving the weaving of the loom, to produce a desired pattern.

- Later, electronic devices were developed that could control mechanical devices with far greater speed, lower energy consumption, and greater distances between information sensing locations and the mechanical control locations.

- Electronic devices also made it easier to use variable temporal delays between acquisition of information and its use for control, and more recently to base control decisions on spatial and temporal patterns in widely distributed sources of information.

Flexible physicalism allows "downward" causation

The pattern in sand cannot change without physical matter being moved.

- Changes in a virtual machine V cannot occur without other changes occurring at the implementation level, in physical machine M.

    That is a consequence of V being implemented in the lower level machine, as explained above.

- So, if changes occur in V (e.g. the pattern visible in the sand) then there must also be changes in M (e.g. the locations of grains of sand) - which will be followed by other physical changes: changes in V are **necessarily** followed by changes in M.

    But the physical routes between the changes in V and in M can vary.

- The cleverness of computer designers (and biological evolution much earlier) included finding combinations of such necessary connections that allow patterns in V to be used to control portions of the material world M: setting up all the links was non-trivial.

- In other words, states, events and processes in a running virtual machine V can not only produce effects in other parts of V, but can also produce effects in the physical machine M in which V is implemented and in some cases can have effects on physical objects, events and processes outside M.

- The resistance to accepting such "downward causation" is based in part on a failure to understand how causal connections relate to a set of factual and counterfactual conditional statements being true.

I am not claiming that "X causes Y" simply **means, by definition,** that a set of counter-factual conditional statements is true: there is also an implied claim that something about the world **makes** those statements true.

Truth-makers can be of different kinds

What sort of thing can make it true that if X had been P, then Y would have been Q?

- Often the connection is far from obvious and finding out why something being P and something being Q are connected in that way is a deep scientific question.

  I think that was also Immanuel Kant's view of causation. ? (?)

- Sometimes the answer depends on physical properties of matter

  – when waves stir up mud, or pushing one end of a rigid pivoted lever down makes the other end go up.

- Sometimes it depends on how information and information-processing systems interact, like a shape in the sand causing you to think of someone.

- Sometimes the relation is mathematical, like inserting a marble into a bottle causing the number of marbles in the bottle to go up, or a bishop move causing a chess piece to be threatened.

- In old machines, primarily concerned with applying energy to move or transform matter (e.g. bulldozers), the "what if?" relation is close to physical laws of nature.

- In some newer machines, like a security system that detects an unrecognized person and immediately locks doors, the key causal mechanisms are information-processing mechanisms. (Here the cause does not provide the energy for the effects.)

- Before the 20th century, most machinery used causation based on transmission of forces, energy and matter, or constraints on such transmission, whereas now our lives are dominated by machines pushing and pulling **information** about! How is that done?

Tangled causal webs

Explaining how a VM works and how the implementation machine makes it work, requires a detailed specification of a great many causal connections.

The causal connections can form a tangled network including:

- causal connections between components and processes in the implementation machine (e.g. between parts of the physical machine)

- causal connections between components and processes in the virtual machine

- causal connections linking components and processes in both machines: linking abstract information patterns and changes in physical processes.

    E.g. allowing signals to cause the printer to start or stop printing, or to change the print quality, or allow a pressed key to alter the contents of a document in a virtual machine.

- causal connections with things outside the physical machine, in the environment.

- Information in a VM can change from being true to being false (or vice versa) without any physical change in the machine. This is semantic, not physical, causation.

    E.g. if Alice is taller than her brother Bob, then Bob's sudden growth can cause Alice to cease being the tallest child in the family, even if they live thousands of miles apart.

    This can alter effects of decisions: e.g. someone who needs to identify Alice after seeing Bob, may misidentify someone else as Alice because he had been told that she was taller than Bob.

    A decision an animal or machine is about to act on may cease to be the right decision because something in the environment has changed, creating an urgent undetected need to acquire more information: causation without physical connection.

Erroneous views about implementation

Sometimes conditions are proposed for supervenience (implementation, realization) of a VM in a PM, that are based on ignorance. E.g. the following must be rejected as *necessary* for supervenience.

- Components of a supervenient system must correspond to fixed physical components which realize them:
  NO: Counter-examples include garbage collection and paging.

- *Types* of VM objects or events must always be implemented in the same *types* of physical objects or events.
  NO: Refuted whenever a running system has its memory replaced piecemeal by newer faster, more reliable, physical components.

- The structural decomposition of a VM (i.e. the part-whole relations) must map onto isomorphic physical structures:
  NO: Counter-examples are circular lists, huge sparse arrays, continuous VM processes implemented discretely, stochastic processes implemented deterministically, ...

- The same temporal relations hold between VM events and physical events as hold between physical events.
  No. E.g. VM time may be discrete while physical events occur in continuous time.

All this means that searching for so-called "neural correlates of mental events requires great conceptual clarity and care.
More likely, it is just misguided: we should be looking for implementations, not correlates.

Causation and counterfactual conditionals in running VMs

A key feature of RVMs in computing systems is not just what they actually do (the internal and external behaviours they actually produce)

but things they would do IF various things happened.

IF a sub-process tries to access a protected file

IF a web page tries to install a trojan horse

IF an inserted character makes a line of text too long

IF a hard drive runs out of space or becomes too fragmented

IF a portion of memory fails when an attempt is made to read or write there.

IF a key is pressed or the mouse is moved

IF a running VM needs to display something, or send a message, or write to physical memory.

A vast amount of engineering effort has gone into producing networks of such causal links in computing systems that

maintain processes within "permitted envelopes"

support required causal interactions

both between VM subsystems and

between hardware and VM subsystems **IN BOTH DIRECTIONS**

E.g. announcing newly arrived email, reporting a memory fault, disabling faulty memory.

Richness of causal networks in RVMs

The network of causal relationships in a modern computing system,

corresponding to all the things that would or would not have happened if something had been different at a particular time

may be vast, yet constantly changing, depending on what happens.

For example at any time there are:
- many different key combinations that might have been pressed but were not;
- many different mouse-actions that might have been performed but were not;
- many different network signals that might have been received, but were not;
- many minor hardware faults that might have been detected and coped with if they had occurred (and some that might have caused particular running programs to crash);
- many different software interrupt triggers that might have occurred but did not;
- thousands of programs that might have been started but were not;
- many running programs that might have run out of space, or might have attempted to access a file system, or might have spawned a sub-process, or might have terminated themselves, but did not;
- many attempts by "malware" programs to violate some restriction, that did not occur but might have, or vice versa;
  ....and many more...

The web of potential causal interactions is too vast to be detected by testing the system from outside to see how it reacts: behaviourism fails for modern computers.

A typical modern PC running windows or linux with typical applications, and an internet connection probably supports tens of thousands of (constantly changing) conditionals, at least.
Or several billions of conditionals if all bits of memory are considered.

Causation and Computer Science

Causation appears to be irrelevant for **mathematical** computer science

A computation can be regarded as just a mathematical structure (possibly infinite), something like a proof.

Such "computations" need not occur in time, nor involve causation:

- A Gödel number encoding a sequence of Gödel numbers can be regarded as a computation: a timeless, static model that accurately reflects all mathematical properties of an actual computation.
- Talking about 'time' in this context is just a matter of talking about position in a (possibly infinite, possibly branching) ordered set.
- State transitions are then not things that happen in time, though they are relations between adjacent components in the ordered mathematical structure.
- The notions of space complexity and time complexity in theoretical computer science refer to purely syntactic properties of a 'trace' of a program execution: another mathematical structure.

Perhaps we should say: theoretical computer science does not study computations, only mathematical models of computations.

Do such models capture important facts about causation, and the possibilities of causal interactions with an environment?

In order to do that, the models would need to allow for branching sets of possibilities, as a Turing machine specification does. (Cf. Kripke semantics for modal logics? 1962??)

Engineering advances enrich the causal networks built and studied (e.g. allowing concurrency, dynamic creation of new concurrent processes, relaxing synchrony, and digital/analog interfaces).

Including physical or human (social) environment in a model assumes we can model physics, and human brains, or minds – but we can't yet!(Sometimes approximations suffice.)

Causation in Computer Applications

People who use computers require more than structural mappings: the machine must be able to **do** things.

There must be causal interactions, happening *reliably* in real time.

Three computers running the same program and voting on results will normally be more reliable than three simulations running on one fast computer even if results are identical when nothing goes wrong.

So, for software engineers, robot designers, and computer users, computation involves a process or collection of processes in which

– things exist
– events and processes happen
– what causes what (e.g. an effect of a bug) can be important

Software engineers want to make some things happen and prevent others

– Some of what happens, or is prevented, is in the virtual machine

(e.g. preventing one user's program from accessing other user's files.)

– Some is in the physical machine (e.g. altering memory or CPU states)
– Some of it is in the environment, under the control of the virtual machine

(e.g. an airliner landing without crashing)

So the engineering notion of implementation/realization goes beyond the mathematical notion of structural mapping. It requires production of causal interactions in the virtual (implemented) machine, in the physical machine, and usually also in the environment.

That includes consumption/dissipation of energy.

Physical time and VM events

Sometimes people want to know exactly when some mental event occurred

E.g. in Libet's experiments `http://en.wikipedia.org/wiki/Benjamin_Libet`

But such questions are often misguided, like asking for the time of an event in a socio-economic virtual machine.

- At which millisecond was the president of the USA elected?
- At which millisecond did the latest economic recession begin?
- At which millisecond did the electorate become disgusted about politicians' expenses claims.

In those cases the question is pointless because the events are statistical events involving the cumulative effect of large numbers of separate events.

But there are other cases: at which millisecond did the 2004 tsunami hit Thailand?

There is no answer because both the coastline and the tsunami were extended in space and different bits made contact at different times.

Also there's no well defined instant at which a wave reaches a pebble on the beach if the wave has an extended cross-section.

The following questions seem to be pointless because the processes involved are extended both in time and spread over different regions of the mind (and brain)

- At which millisecond did I become conscious of my toothache?
- At which millisecond did I decide to go to the dentist?
- At which millisecond did I decide to phone for an appointment?

Before an algorithm finishes running the decision it takes may have become inevitable.

Non-redundant multi-level causation

Some events, including physical events like a valve being shut in an automated chemical plant, may be multiply caused - by physical and by virtual machine events.

NOTE:

This is not like the multiple causation discussed by Pearl and others where removing one of the physical causes would have left a sufficient set of causes

(e.g. being killed by several members of a firing squad, or death by drowning and freezing).

Parallel physical+ VM causation is in an important sense non-redundant: there's no way to simply remove the VM cause of effect E without changing the world in such a way that the physical world ceases to cause E, unlike reducing a firing squad.

Of course the VM cause can be replaced by another: if the bishop had not captured the pawn the rook would have – the physical changes would have had different causes.

There is also the ever-present possibility of total destruction of the whole system, e.g. by a bomb, or a hardware fault that disrupts the virtual machine.

VM counterfactuals depend on a 'normality' condition.

A good implementation tends to preserve normality, e.g. by error detection and error compensation. But there are always limits to what can be prevented: earthquakes, tornadoes, asteroids, bombs...

But that's true also of our normal talk about what causes what. (As Hume noted?)

Virtual Machine events as causes

Most people, including scientists and philosophers in their everyday life, allow causal connections between non-physical events. E.g.

- Ignorance can cause poverty.
- Poverty can cause crime.
- Crime can cause unhappiness.
- Unhappiness can cause a change of government.
- Beliefs and desires can cause decisions, and thereby actions.
- Detecting a threat may cause a chess program to evaluate defensive moves.

How can that be, if all these non-physical phenomena are *fully implemented* in physical phenomena?

For, unless there are causal gaps in physics, there does not seem to be any room for the non-physical events to influence physical processes. This seems to imply that all virtual machines (including minds if they are virtual machines) *must be epiphenomena*.

Some philosophers conclude that if physics has no causal gaps, then human decisions are causally ineffective. Likewise robot decisions.

This is a seriously flawed argument, but exposing the flaws means solving the problem of analysing our concept of causation.

A related discussion of the supposed free will problem can be found in
"How to dispose of the Free Will Issue" (1992)
http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#8

Must non-physical events be epiphenomenal?

Consider a sequence of virtual machine events or states M1, M2, etc. implemented in a physical system with events or states P1, P2, ...
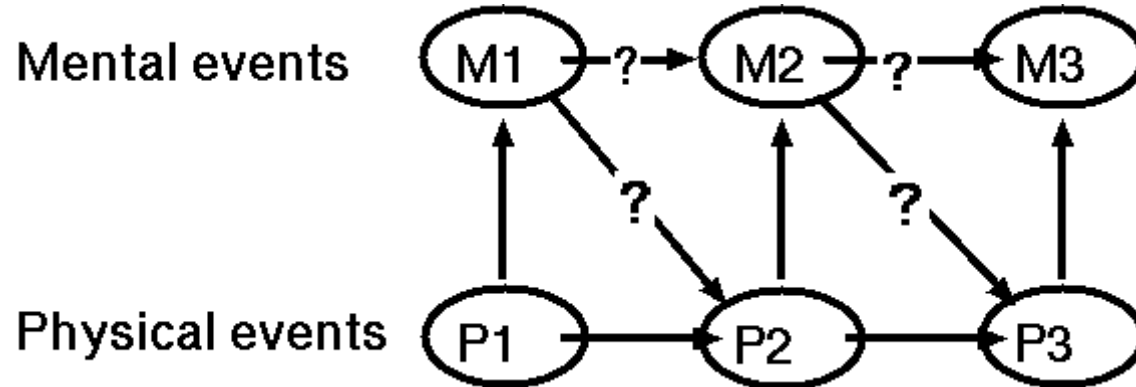
If P2 is caused by its physical precursor, P1, that seems to imply that P2 cannot be caused by M1, and likewise M2 cannot cause P3.

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

So, according to this reasoning

the VM events cannot cause either their physical or their non-physical successors.

E.g. poverty cannot cause broken windows or crime.

This would rule out all the causal relationships represented by arrows with question marks in the diagram, leaving the M events as epiphenomenal.

So the picture must be wrong. In what way?

The flaw in the reasoning?

THIS IS HOW THE ARGUMENT GOES:

IF

- **physical events are physically determined**

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

AND

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

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

THEN

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

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

ONE OF THE CONJUNCTS IN THE ANTECEDENT IS INCORRECT. WHICH?

It's the second conjunct

Some people think the flaw is in the first conjunct:

i.e. they assume that there are some physical events that have no *physical* causes but have some other kind of cause that operates independently of physics, e.g. a spiritual or mental event that has no physical causes.

The real flaw is in the second conjunct:

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

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

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

A sketch of a partial explanation can be found in the last part of this tutorial:

```
http://www.cs.bham.ac.uk/~axs/ijcai01
```

That is expanded here.

NB: Some VM states need external objects

VM events may depend on, be implemented in, "external", even remote, physical events.

- Information in VM X about the external object O can switch from being accurate to being inaccurate simply because O has changed.

- Whether a database is up to date, or complete, is not determined solely by the contents of the physical machine that implements it.

- Reference to particular spatio-temporally located objects requires some external relationship with those objects.

    E.g. for a VM to contain information about the particular individual Julius Caesar, it must have some sort of causal connection with that object, e.g. through informants, or records, etc.
    Otherwise the VM contains only a description of a possible object *similar* to the tower, or to Caesar. ? (?)

- So not ALL mental states of a person or a robot able to relate to an environment are fully implemented *within the body* of that person or robot.
    Supervenience/implementation/realization need not be a "local" relation.

Studying only relations between mind and brain ignoring the physical (and social) environment (methodological solipsism), is a mistake.

(This needs expansion and clarification.)

Biological and artificial VMs

Need to add things about the diversity of VMs produced in evolution.

See the suggestions about evolution of self-monitoring VMs in
`http://www.cs.bham.ac.uk/research/projects/cosy/conferences/mofm-paris-07/sloman`

Virtual Machine functionalism compared with Atomic State functionalism.
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/vm-functionalism.html`

And various talks in

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

To be added
The special case of humans.
Biological virtual machines that
– grow themselves
– learn to monitor themselves
– develop theories about other VMs (in other organisms) etc.
Currently ALL versions of AI are far behind this: embodied, enactivist, dynamical, symbolic, etc.

NOTE:

Most of this presentation was written at various stages prior to Jan 2011.

In the second half of 2011 the Meta-Morphogenesis (M-M) project was born as a result of my reading Turing's 1952 on the Chemical basis of Morphogenesis after being invited to contribute to the (prize winning) book on Turing (?, ?).
Contents listed in `http://www.cs.bham.ac.uk/~axs/amtbook/`

I expect the project to extend significantly the ideas in this presentation.
Further details of the M-M project here (collaborators welcome, but no funds are available at present.):
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html`

Some relevant reading

? (?) includes much that is relevant to the nature of information and causation, though in a very "broad brush" manner. (He is frequently misquoted as defining "information" as "a difference that makes a difference". He actually defines "a unit (or bit) of information" that way. See ? (?). However, that refers to information bearers, not information content.

Some of the analysis of ontological layering in ? (?) is very close to that presented here, except that, like many philosophers, he focuses only on questions about emergence/supervenience of **properties**, doesn't mention computers or running virtual machines, or causation in virtual machines, and does not appear to notice the importance of the omission. Like Jaegwon Kim, he does discuss other cases of causation, e.g. the causal role of temperature in physics.

Some notes on what Dennett says about virtual machines, and his ambivalence about their existence, can be found here: `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk85`

Shanahan's recent book (?, ?) says much of interest about the matters discussed here, especially the discussion of information in Chapter 4, which partly overlaps with the position presented here.

> Chapter 2 gives Wittgenstein's philosophy attention it does not deserve (in this context).
> The real content of the book does not depend on the Wittgensteinian anti-metaphysics.
>
> However, I shall try to show elsewhere that the architecture presented is not adequate to explain the phenomena, partly because it ignores most non-human animals, the differences between humans at different stages of development, and many of the problems evolution had to solve,
>
> > (As discussed in ? (?) and ? (?) and the M-M project proposal: (?, ?).)

Jackie Chappell and I also gave related invited talks on the role of Kantian and Humean causation in human and animal cognition, at a workshop in 2007, Pembroke College Oxford. Details here:

> `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#wonac`
> Causal competences in animals and machines
> > (Including Humean and Kantian causal understanding.)

There are many other relevant publications. Further suggestions welcome.

Some background reading (To be extended)