International Joint Conference on AI – Seattle, August 2001
TUTORIAL SP3
Sunday 5th August 2:00 – 6:00 pm

Philosophical Foundations: Some key questions

Aaron Sloman
and
Matthias Scheutz[1)]

School of Computer Science
The University of Birmingham

http://www.cs.bham.ac.uk/~axs
http://www.nd.edu/~mscheutz

1) Now at University of Notre Dame

Last updated June 3, 2006

The Aim of This Tutorial

This tutorial uses a mixture of lectures and interactive discussions, to introduce philosophical problems relevant to the goals and methodology of:

- **AI as science**

and

- **AI as engineering**

including

- **The contribution of AI to the study of mind.**

Prerequisite knowledge:

Knowledge of AI and experience of software development will help.
Knowledge of philosophy may help or hinder

An overview of the tutorial is available online at

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

This may be expanded after the conference.

AI needs philosophy and philosophy needs AI

Approximate plan

Introduction 20 minutes + 10 minutes questions, etc.

- Introduce main problems
- Summarise some of the main theories.
- 20 slides

Part 2: 40 minutes + up to 20 minutes discussion.

- More detailed analysis of some aspects of the mind/body relation and related topics.
- Functionalism
- Virtual machines
- Supervenience and Implementation,
- About 40 slides

BREAK

Part 3: 80 minutes + up to 20 minutes discussion

- architecture-based concepts of mind
- causal powers of virtual machine events
- Virtual Machine functionalism,
- mechanism supervenience
- About 75 slides

Discussion:

- Final interactive session 30 minutes:

OVERVIEW OF CONTENTS

The tutorial will address the following topics

- Some key concepts and theories in philosophy
- A philosophical view of aspects of AI and software engineering.
  E.g. notions of ‘virtual machine’ and ‘implementation’
- Some key problems in philosophy of mind
  Conceptual problems
  Ontological problems
- Proposals for solving those problems on the basis of
  ‘VIRTUAL MACHINE FUNCTIONALISM’ (VMF)
  THE USE OF ‘ARCHITECTURE-BASED’ CONCEPTS
- Critique of some familiar ideas: e.g.
  The knowledge level
  The intentional stance
  The relevance of Turing machines.

Some Key Concepts and Theories in Philosophy

IMPORTANT CONCEPTS INCLUDE:

- **Epistemology:** the theory of knowledge
- **Metaphysics:** the theory of what exists, and why, etc.
- **Ontology:** an attempt to produce a systematic overview of kinds of things that exist, and their relationships.
- **Philosophy of mind** which includes:
  - **Ontological studies:**
    e.g. what is the relation between mind and matter
  - **Epistemological studies:**
    e.g. how can we know which things have minds - the problem of ‘other minds’
  - **Conceptual studies:**
    What do we mean by various mental concepts, e.g. ‘believe’, ‘desire’, ‘pleasure’, ‘pain’, ‘intend’, ‘perceive’, ‘experience’, ‘consciousness’; what do we mean by ‘cause’, by ‘if ... then’;

    Compare: what do we mean by ‘electron’, ‘energy’, ‘information’ (not in Shannon’s sense).

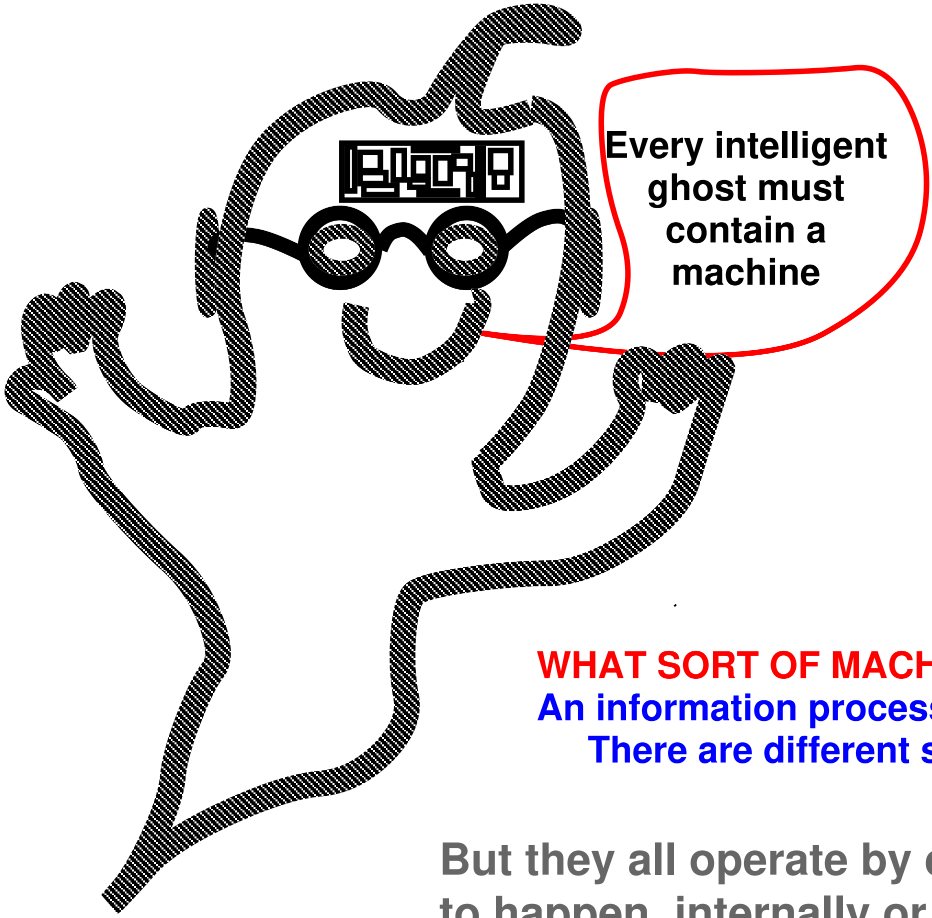
Some Key Concepts and Theories in Philosophy

IMPORTANT THEORIES INCLUDE:

- **Dualism:** mind and matter are two kinds of stuff, each of which can exist without the other
- **Interactionism:** each can causally affect the other
- **Epiphenomenalism:** matter can causally affect mind, but not *vice versa*
- **Preestablished harmony:** neither can affect the other, so any correlation is luck, or a result of pre-set running speeds!
- **Monism:** there is only one kind of stuff, which can be viewed in different ways,
  - **Materialism:** there is nothing but matter
  - **Idealism:** there is nothing but mental stuff
  - **Neutral monism:** mind and matter are both *aspects* of something which is neither mind nor matter.
- **Pluralism:** there are many kinds of stuff, which can exist independently of one another.

Could there be a ghost in the machine?

What philosophers tend to forget - but the ghost of
Gilbert Ryle knows well....



WHAT SORT OF MACHINE?
An information processing machine.
There are different sorts.

But they all operate by causing things
to happen, internally or externally.

Some Key Problems in Philosophy of Mind

Conceptual problems

- How can we define mental states of various kinds?

  E.g. what is their link with behaviour?
- Is desiring a kind of believing (that something will make you happy)?
- What is an emotion?
- What is it to understand (a sentence, a picture, a gesture)?
- What is it to take a decision freely? to be responsible for one's actions?
- Why is it hard to reach agreement on definitions of mental concepts (e.g. emotion, consciousness, believe, learning, decide, etc.)?

  All these are ill-defined "cluster concepts".

  Note that often a conceptual question may have the same form of words as a factual question.

Key Problems .... Continued

Ontological problems

- What kind of existence does a mental state or event have?
- What are mental processes?
  Are they just physical processes viewed in a certain way?
- Can mental events cause (a) other mental events (b) physical events?
- What are the relations between mental processes and physical processes in brains and the environment?
- Puzzles about causation and supervenience. ....
- What is computation? Does it extend our ontology?
  Are computations physical processes or something very different: processes in abstract/virtual machines?

A Philosophical View of Aspects of AI and Software Engineering

- In a working computer the relationship between the running software (the virtual machine) and the underlying hardware is partly like the relation between mental phenomena and physical phenomena. E.g.
  - VM entities, like mental entities, cannot be observed by opening up the physical container.
  - VM entities (e.g. a list structure), like mental entities, do not have physical properties such as mass, volume, charge, etc.
- But there is no deep mystery about events in a Lisp or Prolog VM (e.g. deletion of an item from a list, or unification of a list and a variable), or events in a chess VM (e.g. a capture, or a threat).

  No mystery because we know how to make such things happen.

  We know how to implement, a virtual machine in a physical machine, or in another virtual machine.
- Can that illuminate the relationship between mind and brain?

  We shall discuss this in detail later.

Philosophical views on AI

- **Computational AI is too limited:** Because of the limits of computation, AI systems based on computers cannot even replicate our (e.g. mathematicians') external behaviour.
  (Some of the anti-computationalists promote 'dynamical systems', e.g. van Gelder 1998. For a brief critique see Sloman 1993).
- **The Weak AI thesis:** (Searle 1980) Computer-based robots will be able to *simulate* human capabilities but will not *replicate* mentality.
- **The Strong AI thesis:** Artificial systems will have mental states and processes. This thesis has several forms of varying strength, depending on how the artificial systems are constrained (e.g. discrete/analog, sequential/concurrent, etc.) (Sloman 1992).
- **Computationalism:** A weaker version of the Strong AI thesis, i.e. the claim that mental processes are computational in some sense.
- **The "hard problem":** We may be able to replicate internal and external aspects of human behaviour in every detail, but we have no understanding of *why* this should produce consciousness, sensory qualia (contents of experience) etc. Explaining how these come about is the "hard" problem (Chalmers 1996).

OUR PROPOSAL

We have a proposal that is not generally understood or accepted, and is not found in standard philosophical discussions, though misleadingly similar theories are.

Virtual machine functionalism (VMF):

If there is an appropriate virtual machine whose components interact in the right sorts of ways and have the right sorts of causal powers, the question whether it is really conscious or has real qualia becomes incoherent. This position is elaborated below.

Contrast various kinds of functionalism

- The crude kind many philosophers discuss:
  Consider overall states of the total system associated with input-output mappings.

- The kind software engineers know about:
  A system can have a complex “internal” VM architecture, composed of many coexisting, asynchronously interacting, VM components with their own states

What are virtual machines and how are they related to physical machines?

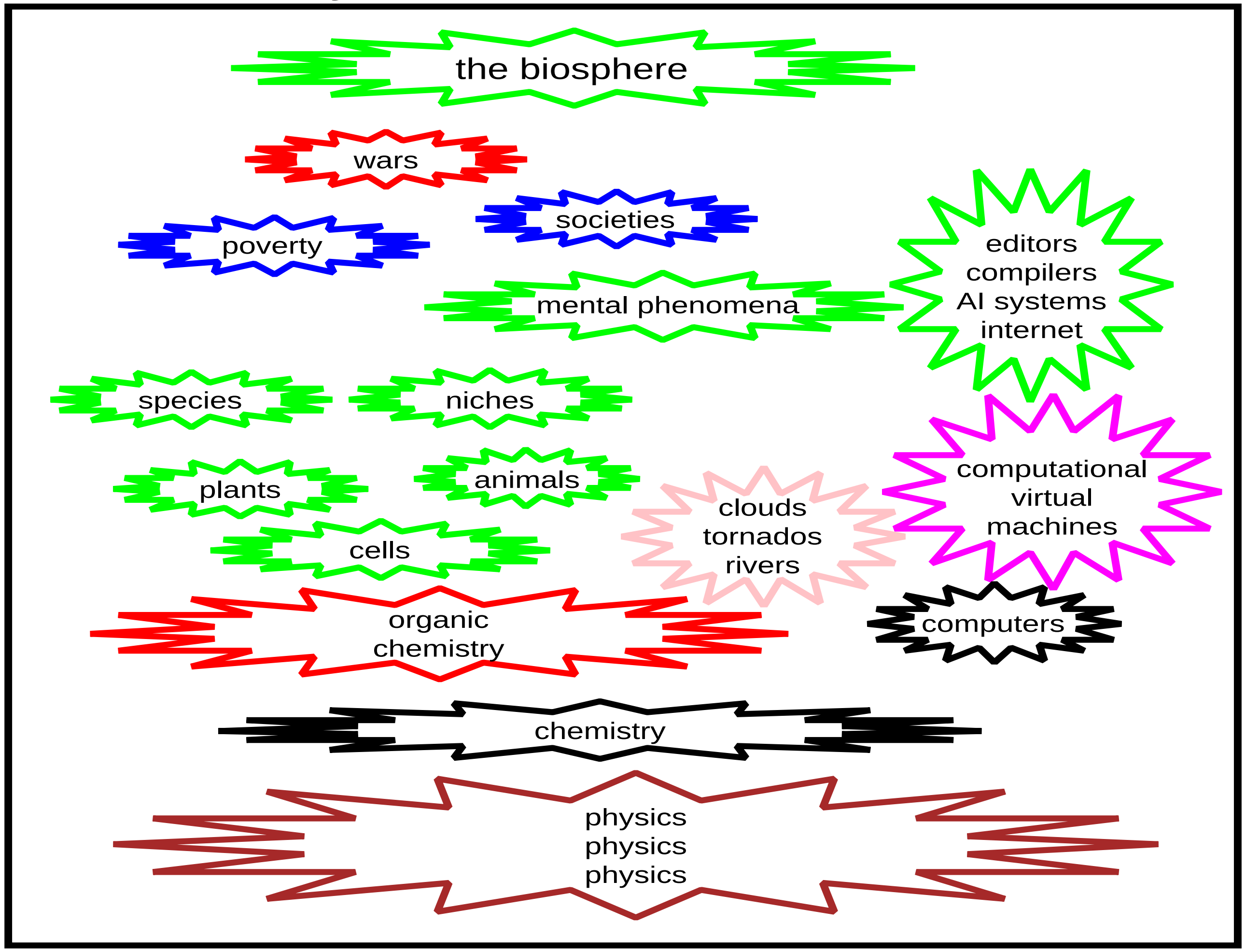
- Can the relations between virtual and physical machines as understood by computer scientists and software engineers shed light on philosophical questions about the relations between minds and brains?
- What are representations, and how many varieties are there in virtual machines?
- What sorts of virtual machines can biological evolution produce?
- What sorts of virtual machine ontologies are required by different organisms or machines?
- What sorts of VM architectures can support ontologies including beliefs, desires, pleasures, pains, decisions, emotions, consciousness, and other mental states, events and processes?
- Can mental events, or events in virtual machines, have causal powers? (Relates to problems about “free-will”).

For more on the nature of information-processing virtual machines, see

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

LEVELS IN REALITY - ONE VIEWPOINT

Multiple levels are everywhere: lots of *REAL* virtual machines:



How many levels of physics?

The importance of architecture

AI researchers used to study mainly algorithms and representations.

Now architectures are seen to be at least as important.

Examples:

SOAR
BDI architectures
ACT, ACT-R, ACT-RP, ACT-RPM
Minsky's *Society of mind* and *Emotion machine.*
Instances of CogAff, H-Cogaff.
'Ecosystem of mind'
Others ....
....

Contrast architectures for:
– single (possibly complex) agents and
– multi-agent systems (often simple, similar agents).

THE MOVE TOWARDS ARCHITECTURES

What does that mean?

Why has there been such a shift?

- **Engineering considerations:**
  different ways of decomposing complex systems.
- **Scientific considerations:**
  organisms do not work like typical programming language VM (i.e. single thread of control with a hierarchy of procedure activations).
- **The importance of concurrency and asynchrony:**
  coexisting, asynchronously interacting components (including the environment and transducers): not like Turing machines.
- We hardly understand the space of possible designs.
- So beware of dogmatic claims
  - Everything is a neural net
  - Everything is a dynamical system
  - Everything is a collection of reactive behaviours.
  - Turing machines can do everything

Architectures vs Algorithms & Representations

Thinking about architectures can change some old discussions.

E.g. there is much discussion of a distinction between procedural and declarative representations, as if this were primarily a distinction of syntactic form. (See Sun *et al.* 2001.)

We can instead talk about

the location of the representation within an architecture: which mechanisms can access it and what they can do with it.

E.g. even the most "procedural" of representations in computers, compiled machine-code procedures can be treated as "declarative" by mechanisms which create them, analyse them relocate them, optimise them, etc, instead of simply *running* them.

Similar points could be made about alleged differences between symbolic and subsymbolic representations: for us this will turn out to be more a matter of whether the representations are used only by *reactive* mechanisms or also by *deliberative* mechanisms (explained later).

Likewise the distinction between verbal (logical, Fregean) and pictorial (analogical, spatial, iconic) representations. (Sloman 1996a)

Mechanism supervenience

The architectures we are interested in are

Virtual machine architectures.

We are very familiar with virtual machines, and can build them and use them.

However the kinds we are familiar with are probably a tiny subset of what is possible.

Analysing the nature of the ones we know, and explaining how they relate to physical machines is not easy: that's a philosophical problem, on which there is much disagreement.

Questions:

- Is there an interesting subclass of VMs that could be produced by processes of biological evolution?
- What sorts can and cannot be so produced?

(Compare the notion of "trajectories in design space and niche space" (Sloman 2000))

Some key ideas about Virtual Machines

Virtual machines are typically structure-manipulating machines.

They may be continuous (analog) or discrete (digital) or fuzzy.

They may have one serial processing stream or many parts operating concurrently.

They depend on and are implemented in ‘lower level’ machines: which may also be VMs or may be physical machines.

They operate on abstract structures, e.g. lists, trees, numbers.

Their parts need not map onto physical components in any simple way.

Events in VMs can cause other events, in the same VM, in a physical machine, or in another VM.

The relationship between a virtual machine and the implementing machine involves whole ontologies containing many interacting components .

SUMMARY OF CLAIMS SO FAR

- Old philosophical concepts and theories can be illuminated in the light of what we have learnt through the development of computing
- In particular many of the problems that philosophers have discussed about the nature of mind and the relation between mind and body can be clarified by relating them to what we have learnt about the nature of virtual machines and the relationships between virtual and physical machines.
- What we have learnt about the latter is often understood only intuitively by people who develop computing systems: making that knowledge explicit is not easy.
- This is a form of dualism but not “substance dualism” – minds cannot exist independent of matter.

End of part 1

A short time for questions and discussion

PART TWO: Towards Virtual Machine Functionalism...

- Problems of mind and body
- How are mind and body related?
- Virtual machines
- Supervenience and implementation

Mind and Body

- What is the relation between the mind and the body?
- This is called the mind-body problem–why is this a problem?
- Very old, still unresolved problem in philosophy (has become very important especially in the 20th century)
- Before being able to provide an answer to the mind-body problem, we need to be clear on what mind and body are: the same or different kinds of things, i.e., one or two *substances*
- What do you think? (E.g., what do Christians think? What do Star Trek fans think – How could a physical transporter suffice to transport your mind?)
- Problem with two substances: how can there be interaction between them? (Why is this a problem–does there have to be any interaction?)
- Problem with one substance: why do mental phenomena seem to be so different from bodily or physical phenomena? (Why is this a problem–does the mental have to be different?)

Two Metaphysical Positions on Mind and Matter

- **Reductionism:**
  All *mental properties* (such as having emotions, intentions, desires, plans, feelings, perceptions, etc.) are *physical properties* (in the sense of the most comprehensive available physical theory) and can be ultimately defined in terms of them (e.g., having emotions could be seen as "nothing but" having certain hormonal concentrations in the blood stream)

- **Antireductionism - negative answer:**
  no, that is not possible, because ....!

- **Antireductionism - positive answer:**
  while the mental might not be *reducible* to the physical, it is nevertheless *determined by* and *dependent on* the physical
- In other words, the the mental *supervenes* on the physical without being identical to it (i.e., without being necessarily reducible)
- Will get back to mind-body supervenience later

Main Approaches towards the Relation of Mind and Matter

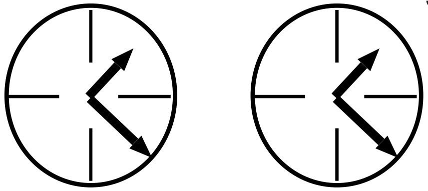
- Distinguish two dimensions
  - Substance dualism - substance monism
  - Property dualism - property monism
- The four possible stances regarding the nature of mind and matter:
  - Substance dualism + property dualism
    usually called “Cartesian dualism” (two substances with their own sets of properties)
  - Substance dualism + property monism
    seems to make no sense (what could it mean to have two different substances that are characterised by the same set of properties?)
  - Substance monism + property monism
    usually called “reductive physicalism” (only one substance, usually matter, and its properties)
  - Substance monism + property dualism
    usually called “irreductive physicalism” (only one substance, but different sets of properties, that characterise it)

Mind and Body: Substance Dualists

Remember: every substance dualist is also a property dualist

All of the following positions have been largely abandoned in the philosophy of mind:

- **Causal interactionism**
  (Descartes, causal interaction between the mental and the physical via pineal gland)
- **Occasionalism**
  (Malebranche, no direct interaction between mental and physical at all, God is the mediator who creates it)
- **"Preestablished harmony"**
  between mind and body (Leibniz, no interaction, rather everything is set up by God in advance so that the right kinds of events happen in the mental and physical at the right time, just "as if" they interacted, like two clocks telling the same time.)



Mind and Body: Substance Monists

The following are substance monist accounts, yet none of these positions is very popular anymore (for various, partly different reasons):

- **The double-aspect theory ('neutral monism')**
  (Spinoza, mind and body are two distinguishable aspects of the same substance)
- **Epiphenomenalism**
  (Huxley, every mental event is caused by a physical event, but mental events have no power of their own)
- **Emergentism**
  (Alexander, mental phenomena emerge from complex physical aggregates and this fact cannot be explained)
- **(German) Idealism:**
  ideas (i.e., the mental) are the real substances (Berkeley, Schelling, Fichte, Hegel and other German idealists)

Mind and Body: Materialism and Physicalism

Nowadays most philosophers and philosophical scientists are substance monists (while people in the neurosciences are usually property monists, people in cognitive science and AI are usually property dualists)

- Materialism/physicalism:
  Matter is the real substance and physics is going to tell us what it is
- Different kinds of behaviourism:
  Philosophical (e.g., logical, ontological - Hempel, Ryle), Psychological (e.g., methodological - Watson, Skinner)
- Identity theory:
  Mental *states* (or properties) are physical *states* (or properties)
  - Identity theories come in two versions depending on whether the term “state” is interpreted as *state type* or *state token*
  - Roughly, a “token” is a concrete instantiation of a type (e.g., there are many tokens of the letter ‘e’ on this slide)

Mind and Body Identity Theories

- Hence, the claims made by identity theorists can be classified accordingly:
  - mental *state tokens* are physical *state tokens* (every monist believes that)
  - mental *state types* are physical *state types* (many reductionists believe that, e.g., Armstrong, Smart, Feigl, *et al.*: brain and mind states are identical, reductionism!)
- Note that token identity is a very weak claim: it effectively says that every time a mental property is instantiated, a physical property is also instantiated such that the instance of the mental property *is* the instance of the physical property
- Example: Suppose that C-fibres are brain mechanisms that can be responsible for pain in humans. Then according to the *token identity theory* it is true that if I have a C-fibre activation at time t then I also have a pain at time t.
- However, it is not necessary that every time I have a pain, I also have a C-fibre activation (e.g., I could have a D-fibre activation instead)

Problems with Type Reductionism (multiple realisability)

The mind-body **type identity** claim, on the other hand, seems too strong:

- Suppose that C-fibres are the brain mechanisms responsible for pain *in humans*, then according to the type identity theory we might be tempted to claim the identity

  $pain :=_{def} C\text{-}fibre\ activation$!

- But what about Martians, who also have pain, but no C-fibres (instead they may have D-fibres)?

- Basic problem: mental states can be *multiply realised* (e.g., pain in C- or D-fibres), hence they cannot be "identical" to physical states (i.e., either C- or D-fibre activation).
- Pains might still be identical to C- or D-fibre activations, but then we need to make clear whether and why this disjunction is a physical property (there is a whole discussion of whether a disjunction of physical properties is or is not a physical property)

Eliminative Materialism

- One way to stay reductionist and not battle with those kinds of issues is "eliminative materialism", which states:

  Folk psychology (and all other sciences making use of concepts such as "pains", "beliefs", "hopes", "emotions", etc.) is misguided–there really are no such things!

  – Do we want to be so radical?
  – Are these concepts all "pre-scientific"?
  – Do they all need to be put in the conceptual garbage bin after our failure to make them precise?

- Later we will see how families of architecture-based concepts can extend and refine our pre-scientific concepts, rather than eliminating them.

  Compare what a new theory of the architecture of matter did for our concepts of "carbon", "iron", "water", "salt", etc.

Functionalism

- **Functionalism** combines advantages of identity theories (that the mental is related to the physical) and behaviourism (that input-output behaviour matters): *mental states* are *functional states*
- A *functional state* is individuated with respect to its *function*, i.e., the role it plays in the overall *functional architecture* of the system
- Two main variants: define functional states

  EITHER in terms of

  (a) their causal role (usually with respect to producing external physical behaviour based on given physical inputs)

  OR in terms of

  (b) states in a *state transitions table* (similar to the control states of a *Turing machine* or *finite state automaton*, say)
- Functional states are then "physically realised" in the system that "realises" the functional architecture, where "functional architecture" usually simply denotes the whole set of functional states and their transition relation.

Mind and Body: The Advantages of Functionalism

- **Functionalism avoids two major problems:**

  (1) the identification of mental and physical properties of type identity theory

  (2) the inability of behaviourism to speak about "inner states" (which can explain behavioural dispositions)

Functionalism and Behaviourism

- **Main differences between standard functionalism and behaviourism:**
  - Functionalism talks about **real inner states** of an organism with “causal powers”, which behaviourism was strongly opposed to.
  - I.e., functionalists are **realists** about dispositions, behaviourists are **instrumentalists** – (I/O must be entirely “observable” for both, but in addition the functionalist allows reference to mental states, which are defined in terms of input, output, and other mental states)

- **NOTE:**

  “Virtual Machine functionalism”, described later, and in

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

  goes further and allows internal mechanisms to be defined *entirely* in terms of their dispositions to interact with other internal mechanisms: they need not be modifiable by external events or produce external effects.

  However, many of them will be related to external inputs and outputs, though often very indirectly.

Functionalism and identity theories

- Main difference between functionalism and identity theories:
  in most versions of functionalism,
  – mental states are identified with functional states,
  however
  – they can be multiply realised in the physical,
  hence
  – type identity is abandoned in favor of mere token identity (functionalism has been criticised as too “non-committal” for that reason!)
- Later we’ll query even token identity.

Block's Example

- What does it mean for a system to realise a "functional architecture" (i.e., a set of functional states (of type (a)), which are mutually defined in terms of their causal roles)?
- Put differently: given a physical system *S* and a functional architecture *FA*, when can we say that *S* realises *FA*?
- A functionalist account of what it means to be in state *E* (of the "even-odd" automaton, where '*E*' and '*O*' stand for "even" and "odd") would look like this:

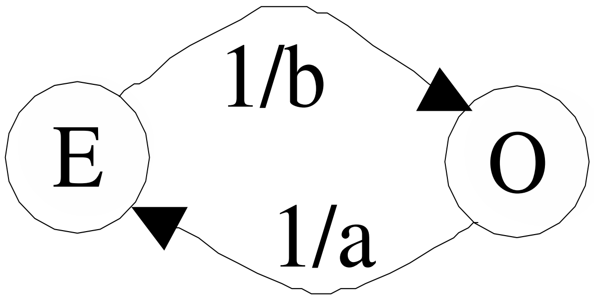
being in $E =_{def}$

being an x such that $\exists P, Q$

(x is in P

$\wedge$ (if x is in P and gets input '1', then it goes into Q and outputs 'b')

$\wedge$ (if x is in Q and gets input '1', then it goes into P and outputs 'a'))



Functional States: Example Cont'd

- **Note that implicit in the definition of functional states is the idea that functional states are realised physically in that they *are* physical states–why?**
- **Because we do not "add" anything by quantifying over first-order (i.e., physical) properties:**

  for something to have E is to have one of its realisers P
  (i.e., one of the Ps that satisfies the right hand side of the definition of E)

  being in $E =_{def}$

  being an x such that $\exists P, Q$

  (x is in P

  $\wedge$ (if x is in P and gets input '1', then it goes into Q and outputs 'b')

  $\wedge$ (if x is in Q and gets input '1', then it goes into P and outputs 'a'))
- **Note: the existential quantifier in the definition of functional states leaves open whether there is more than one realiser ("multiple realization") – there could be a P_1 and a P_2, both of which realise E in different systems**

Problems with Functional States

- **The previous definition of functional states, however, seems to conflate functional states and their realization: it is not possible to identify two functional states in two *different* "functional architectures" (why?)**
- **Furthermore, it is not clear why input/output states should not also be functional (e.g., if they can be realised in many ways as in the previous case of '1') – functionalists usually take inputs and outputs as physically specified (Chomsky 1959 criticises Skinner on this count)**
- **Functional states are too *coarse-grained* to be useful in explanations of the functionality of the system as they are states of the whole system (there are many problems involving *time*, *extendibility of the architecture*, *redundancy of the functional description*, *states vs. processes*, etc.**
- **Later we shall introduce a particular form of functionalism based on the notion of "virtual machine", which views mental states as intrinsically architecture-dependent.**

Mind and Body: Variations of Functionalism

- **Cognitivism: need to account for the complexity and representational structure of the mind ("intelligent behaviour can only be explained by appeal to 'internal cognitive processes' ", J. Haugeland, *Mind Design I*)**
- **Representationalism: minds create and manipulate representations (though not necessarily Fodor's "language of thought")**
- **Computationalism (dominant view in cognitive science since its beginnings): mental states are computational states.**
- **Motivating idea: minds process information, computers are the paradigmatic information processors, hence mental processes may be (at least partly) computational processes and thus (partly) describable in terms of programs.**
  (Later we return to the question "What is computation?".)
- **Note that computationalism was the driving force behind AI and cognitive science from their beginning (but has also been criticised ever since)**

Questions about “Virtual Machines”

- The term “virtual machine” is used very frequently in computer science (e.g., the Java Virtual Machine JVM)
- But what exactly do people mean by “virtual machine” (VM)? Why “virtual”?
- What do VMs do?
- Are VMs unreal?
- How do VMs differ from physical machines?
- How do VMs relate to physical machines?
- How can VMs be produced?
- Why are VMs relevant to AI and to cognitive science?

Machines and Architectures

- Machines are generally complex systems with an architecture, which is defined by the number and kinds of interacting components and their contributions to the capabilities of the (overall) machine
- An *architecture*, therefore, specifies a particular *type* of machine (by fixing the components and their interplay) without going into detail about the physical nature of the components (e.g., what the components are made out)
- Architectures are blueprints that help engineers construct and build machines (e.g., think of the circuit diagrams)
- Note that architectures as schematic ways of specifying machines depend to a large extent on pragmatic information about what the components are and how they can be realised physically to be meaningful (e.g., a circuit diagram might mention different types of transistors with particular functional specifications)
- Note that components of machines may also be machines themselves with their own architectures (although in every specification there are smallest components–no infinite regress!)

Machines and Components

- Often the architecture of a physical machine involves a special physical decomposition of the overall physical system into spatially distinct physical parts that serve distinct functional roles
- Components of machines themselves are typically specified functionally: what they are used for and how they are used is part of the description of what they are (e.g., spark plug)
- To see that parts of a machine are usually functionally specified think of decomposing a car into parts: just cutting it in half with a chain saw will produce two "physical parts", but without any functional role
- Sometimes these functional definitions contain mathematical descriptions of the part's geometry (e.g., "thread") or other details that are required for any physical instantiation of it
- In addition to the functional parts, components might also have physical parts in their definition (e.g., a "power cable" needs to be a conductor for it to be a power cable, hence it needs to have specific physical properties, which are usually instantiated by metals)

Virtual Machines

- **Programs are very similar to functional specifications of architectures of machines in that they also specify parts (e.g., data structures) and how they are related**
  - **spatially (e.g., where in (virtual) memory, or in abstract data-structures, they are placed) and**
  - **temporally (e.g., how they are created and can change over time) through the specification of processes that operate on them**
  - **causally (e.g. which procedures run when, and what happens)**
- **Hence, we can say that software engineers and programmers also build machines, not physical machines, though, but rather virtual machines (VMs), which are *information processing machines***
- **It is important to understand that while VMs are not physical machines, this does not make them any more real or unreal than physical machines as they are (eventually) *implemented* in physical machines (e.g., computers)**
- **Familiar examples of VMs include chess playing programs, word processors, email systems, operating systems, compilers, the SCHEME virtual machine**

Virtual Machines and Physical Machines

- **While VMs are *implemented* in terms of lower level (ultimately physical) machines, the main features of virtual machines are *not* physical properties and physical behaviours**
- **Rather than physical objects and their physical properties, virtual machines manipulate complex information structures**
- **Claim: Interacting components of a virtual machine do not interact *via* physical causes (such as forces or voltages) even though they are implemented in terms of machines that do.**
- **Just think of a PC, where a virus *causes* the disk drive to be erased – the above claim effectively states that it is the virus, which causes the disk to be erased rather than the sequence of changes in the strength of the electromagnetic field at the set of spatial locations, which instantiate the PC.**
- **This is a strong, very condensed statement that needs to be spelled out in detail**

  For a hardware engineer it may be the physical phenomenon that is best described as the cause. Later we explain why these views are not inconsistent.

Virtual Machines and Philosophy

- The notion of "causation" plays a central role in our understanding and explanation of the behaviours of VMs (hence, "causation" needs to be explicated–will get to this in more detail later)
- Also: compare the previous claim to the problem of mental causation in the philosophy of mind
- Consider the analogy mind : matter = VM : matter (often called "computer metaphor", though that could be misleading if there are non-computational VMs)
- Idea: philosophy and computer science could instruct each other!
- Another interesting philosophical feature of VMs is that they can have their own set of concepts, which define and are defined for the realm (and level of abstraction/description) in/on which they work. This is often referred to as the ontology of the virtual machine.
- In chess, for example, the concepts include "threat", "check-mate", "knight move", etc., whereas in email systems the concepts are "addressee", "message body", "attachments", etc. So a chess VM and an email handling VM will use different ontologies.

An Example of a Virtual Machine: A Word Processor

- Said before that VMs have their own set of concepts, and hence any description of a VM will make use of properties, functions, relations, etc. particular to the domain, on which the VM operates
- In philosophical jargon: VMs have their ontology!
- A word-processor, for example, manipulates words, paragraphs, etc., though these cannot be found in the underlying physical machine.
- More specifically, the architecture of a word processor virtual machine uses an ontology that includes
  - entities like letters, words, paragraphs, etc.
  - properties like italics, bold, double-spaced, etc.
  - relations like in-between, first-on-page, anchored-at, etc.
- Nowhere in the ontology of a word processing virtual machine are bits, bytes, registers, for-loops, stacks, arrays, procedures, etc. mentioned–rather they are part of the "implementing" virtual machine

Virtual machines and Implementation

- How can we implement virtual machines if their ontology can be different from the one of the implementing system? (Obviously, it can be done, since we do have implementations of chess computers, word processors, etc.)
- Computer scientists, and particularly engineers, have strong intuitions about what it means to implement virtual machines
- Yet, many concepts used to describe the various implementation relationships are not made precise and explicit enough to cover all the interesting aspects of implementations (most of which might be relevant to the mind-body relation as well)
- The possible relationships between virtual machines and their implementing systems can be very complex and hard to understand
- It is very difficult to keep the following two facts conceptually separate: (1) that VMs can be defined in terms of concepts, which are different from typical physical ones, and (2) that VMs get their causal power, i.e., do work because they are implemented in physical systems–this is the "implementation problem of VMs" and we will talk about it later

Virtual machines and Information

- Many VMs, as information processing machines, do not have a meaningful physical description (e.g., what would it mean to describe “chess” in terms of physics?)
- So, while a PC may implement a chess program, it will be extremely hard if not impossible to understand what it does if one looks at the level of machine code or electric fields alone: to understand, what the “physical system PC” is doing, one needs to understand it at the level of the “virtual chess machine” it is implementing
- In general, since functions in a virtual machine, such as deriving new information, making a plan, choosing a chess move, or multiplying two numbers are *not* physical functions, it will be futile in most (interesting) cases to look at the implementation level to gain insights into what they system does (e.g., because any given system implements infinitely many VMs–why is this?)
- At the VM level, however, we will be able to understand its behaviour (but note that without the architecture of the system, it is not clear what *the VM level* is supposed to be)

Why is this important?

- Because of the following claim:

  In "intelligent control systems" most of the important processes (of the controller) are likely to be found in such virtual machines.

- Since they are implemented in physical machinery (whether computers or brains) there will of course be processes involving transistors or neurons and certainly atoms and sub-atomic particles and fields.
- However, describing what is going on at the physical level will not provide any information at a useful level of generality.

  For instance it will not be possible to make useful comparisons between different control processes implemented on different physical hardware, where the difference may be that one uses a depth first search for a good decision while another searches breadth first, and a third attempts to use previously learnt stored patterns and searches only if it cannot find one.

  Virtual machines that use the same decision strategy may be implemented in very different physical machines – multiple realisability again.

Supervenience

- Remember: we used the notion of "supervenience" in connection with the positive version of the antireductionist claim about the mind-body relation, that the *mental supervenes on the physical* (i.e., that the mental is determined by and dependent on the physical, but not reducible to it)
- Originally, it was used by G.E. Moore to characterise the relationship between ethical and factual statements:
  - if two actions or situations are exactly similar in all their "natural" (non-moral) features then they could not differ in goodness or badness
- Later, Davidson (1970) introduced the notion of supervenience into the mind-body debate:
  - "[...] supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respects, or that an object cannot alter in some mental respect without altering in some physical respect." (1980, p. 214)

Notions of Supervenience

- At least three different notions of supervenience (all only defined for properties):
  - strong
  - weak
  - global
- Whereas weak and strong supervenience apply to properties of individuals or persons, global supervenience applies to properties of whole worlds
- Other supervenience notions:
  - pattern supervenience (e.g., the supervenience of a rotating square on the pixel matrix of a computer screen)
  - mereological supervenience (e.g., the supervenience of a pile on grains of sand)
  - mechanism supervenience (e.g., the supervenience of a WORD processor on the PC hardware–this is the one we are interested in for intelligent control systems)

Formal Definitions of Supervenience

To define the three most common notions of supervenience, fix two sets of properties α and β. Then we say that

α-properties *supervene* on β-properties
iff

- **(weak) necessarily, for any mental property α:
if anything has α, there exists a property β s.t. it has β, and anything that has β has α.**
- **(strong) necessarily, for any mental property α:
if anything has α, there exists a property β s.t. it has β, and necessarily, anything that has β has α.**
- **(Global) For any worlds w_i and w_j,
if w_i and w_j are β-indiscernible, then w_i and w_j are α-indiscernible.**

Supervenience and the Mind-Body Problem

- Is supervenience an account of the mind-body relation (i.e., a solution to the mind-body problem)?
- No, because it is compatible with many mutually exclusive mind-body theories (e.g., type identity theory, epiphenomenalism, emergentism, etc.)
- In part that's because those notions of supervenience say nothing about *causation*, e.g. whether a supervening property can have *causal effects*.

  "We must conclude then that mind-body supervenience itself is not an *explanatory theory*; it merely states a pattern of property covariation between the mental and the physical and points to the existence of a dependency relation between the two. [...] mind-body supervenience *states* the mind-body problem-it is not a solution to it." Kim (1998, p. 14)

However there may be a way of thinking about supervenience that interprets "A supervenes on B" as very close in meaning to "A (considered as a VM) is *implemented* in B, or *realised as* B".

The Physical Realization Thesis

- **If the most common notion of supervenience only states the mind-body problem, what can help us explain the relation between mind and body?**
- **Suggestion: the notion of physical realization**
- **The Physical Realization Thesis (PRT): *the mental is realised in the physical* (or put differently, if mental properties/states are realised, they have to be physically realised).**
- **As it stands, PRT is too contracted, needs to be disentangled**
- **Suggestion: take a look at the notion of implementation as used in computer science**
- **Many of problems connected to functional architectures and their realization arise in the context of implementation of programs as well**
- **In particular, the closest parallel would be the notion of "implementing a virtual machine in a physical machine"**

NOTE: the physical realisation thesis is probably what Newell and Simon meant by their "physical symbol system" hypotheses.

The Intuitive View of Implementation

- The intuitive view of implementation in computer science is that there exists some sort of *correspondence* between physical and computational states
- Open issue: what *exactly* are the properties of this correspondence?
- E.g., how do data structures in C correspond to bits and bytes? Or how do definitions of computational processes in C correspond to definitions of computational processes in assembly?
- Usually, “interpreters” and “compilers” need to solve this problem for particular instances, i.e., for particular high-level and low-level programming languages!
- Note that this view depends on the notion of *physical states* (assuming that it is clear what computational states are), i.e., it is *parameterised* by a *physical theory* (which delivers the notion of physical state)

Problems with the Intuitive (Correspondence) View on Implementation

- Among others, there are two well-known arguments raised by philosophers against the "intuitive notion of implementation":
  - Searle's argument (sketch) that walls *implement* the WORDSTAR program
  - Putnam's argument that any open physical system *realises* every finite state automaton (without input and output)
- Following Searle's and Putnam's definitions, there are two views on implementation:
  - The semantic view SV (*implementation* as the *interpretation relation between certain formal theories and models thereof*)
  - The state-to-state correspondence view CV (*implementation* as the *correspondence relation between computational and physical states*)

Implementation of Functional Architectures

- Definitions of "implementation" have been offered for both views, which are supposed to block the unwanted implementation results (yet, these philosophical definitions of implementation seem only remotely related to practical notions as understood and used by software engineers, e.g., when they implement an operating system)
- Main problem: there are cases (as with VMs), when we are interested in details of the internal structure (e.g., in parts of the architecture and how they work together), but philosophical notions usually only deal with overall states of the system
- This was the main problem of the standard version of functionalism (as discussed in Block's example): functional states are too coarse-grained to capture internal structure–they only capture the "overall state" of the system
- Need to be able to talk about parts and their interconnections: how should we group them to obtain functional units?
- Difficulty: the same physical part can participate in different functional units

Implementation of VMs

- For VMs this can be done (in simple cases) by relating entities of the VM to (sets of) parts in the physical system in such a way that this implementation relation preserves transitions in the VM and its implementing system (and in more complex cases using relations between relations)
- However, this relation may be "partial" (if considered at a particular point in time) in that there can be entities in the VM which do not correspond to parts at that time
- Example: A linear algebra VM, which contains huge sparse arrays of various elements as part of its architecture, could be implemented on a computer with fewer parts (i.e., memory cells) than there are items in the array by using compressing techniques (e.g., 0 items are not explicitly represented)
- VM entities may correspond directly to low-level entities, or they might correspond to complex low-level entities (or vice versa)
- This correspondence does not have to be same for the life time of the VM (e.g., think of a virtual memory system, or a garbage collector, where mappings between memory locations between virtual and physical memory can change over time.)

Implementation of VMs and Supervenience

- Main difficulty with the implementation of a VM: two ontologies may have to be related when implementing a VM–the VM ontology and the ontology of the implementing system–without necessarily *reducing* one to the other
- For example, a word processor, whose ontology contains
  - entities like letters, words, paragraphs, etc.,
  - properties like italics, bold, double-spaced, etc. and
  - relations like in-between, first-on-page, etc.

  may be implemented on a PC, whose ontology contains
  - entities like bits, bytes, addresses, registers, etc.,
  - properties like even, odd, n-th-in-a-byte, etc. and
  - relations like content-of-memory-location, =, etc.
- This problem is extremely difficult to tackle, yet it seems to come up in various contexts in AI

Implementation/supervenience without reduction

- Once a VM is implemented in a system, however, the VM *strongly supervenes* on the implementing system
- So implementation is an answer to the “mind-body problem for VMs”:

  The VM is determined by and dependent on the implementing system – without the VM ontology being reduced to the ontology of the implementing system!
- When a VM is implemented in a machine whose ontology is P (e.g. physics), people are sometimes tempted to say the VM is “nothing but” P, e.g., thoughts are *nothing but* atomic and molecular processes in brains, a process of checking the spelling in a document is *nothing but* a collection of electronic processes in transistors. (the “nothing buttery” fallacy).
- This is grossly misleading when the ontology of VM is quite different from that of P. Is the spread of information about a declaration of war nothing but electromagnetic radiation and movements of atoms in ink, and paper and brains?

Non-features of mechanism supervenience

Sometimes conditions are proposed for supervenience that are violated by examples from computing. E.g. the following must be rejected as *necessary* for supervenience.

- **Components of a supervenient system must correspond to fixed physical components which realise them:**
  Counter-examples were mentioned above.
- ***Types* of VM objects or events must always be implemented in the same *types* of physical objects or events.**
  Refuted by the recent history of computing.
- **The structural decomposition of a VM (i.e. the part-whole relations) must map onto isomorphic physical structures: NO.**
  However, list A can be an element in list B while B is an element in list A, whereas no two physical objects can contain each other as parts. Also sparse arrays, etc.
- **If a VM, M, is implemented in a physical machine, P, then P must have at least as many physical components as M:**
  No, for the same reasons.

Does all this mean that searching for so-called “neural correlates of consciousness” is misguided?

Some features of mechanism supervenience

- For a *working* instance of a type of VM to exist there must be some working physical mechanism that implements it. I.e. virtual machines depend on physical systems.

  By contrast theoretical VMs can be studied without presuming that any actual running version exists, or could exist.

- A VM difference is impossible without some physical difference.

  G.E.Moore required this for ethical properties supervening on non-ethical properties.

  – If a running VM changes in some way there must have been a physical change.
  – If two VMs M1 and M2 differ in some VM feature there must also be a physical difference in the implementing machines or their environments.

    NOTE: *Difference in physical machines does not imply difference in VMs, but difference in VMs implies physical differences.*

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 Y can switch from being accurate to being inaccurate simply because Y 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 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. (Strawson 1959)
- **So not ALL mental states of a person or a robot are fully implemented *within the body* of that person or robot. Supervenience need not be a “local” relation.**

  Trying to study only the relation between mind and brain, ignoring the physical (and social) environment, is a mistake.

SUMMARY: Important topics discussed in this section

virtual machine functionalism

supervenience

implementation

Have we solved the mind body problem?

PART THREE
ARCHITECTURE BASED CONCEPTS
AND THE COGAFF FRAMEWORK

- **Many of our mental concepts are muddled**
- **Partial diagnosis: blind men feeling an elephant**
- **Partial diagnosis: many of our concepts are ill-defined “cluster concepts”**
- **Partial solution: architecture-based concepts can be made far more precise**
- **Further problem: there are many archictures: so use an evolutionary perspective.**
- **COGAFF: a schema for thinking about variety of evolvable architectures, and some others too**
- **A special case: H-COGAFF — A conjectured human-like architecture**

Many of our mental concepts are muddled

- **Partial diagnosis: we know about only a small part of the subspace**
- **Remedy: develop architecture-based concepts for multiple virtual machine architectures.**
- **The CogAff architecture schema described and the H-Cogaff, human special case, sketched.**
- **Some implications, and work to be done**

Architecture-based concepts of mind

- **Problem:** how do our concepts of mind work? We don't know, but we think we know; our ideas are very confused.
- **Partial Diagnosis:** We are like the ten blind men trying to describe an elephant. The elephant is the space of possible mental concepts.
- **What to do:** depends on motivation for studying mental concepts. Possible motives include:

  (i) **Science:** An interest in natural emotions (in humans and other animals) as something to be modelled and explained, or an investigation of how they might have evolved, etc.

  (ii) **Improved interaction:** A desire to give machines which have to interact with humans an understanding of emotions as a requirement for some aspects of that task.

  (iii) **Entertainment:** A desire to produce new kinds of computer-based entertainments where synthetic agents, e.g. software agents or "toy" robots, produce convincing emotional behaviour.

  (iv) **Educational tools:** E.g. building models for psychology students to play with

  (v) **Therapeutic aims:** Using models in diagnosis and design of therapies.

Muddles about mental concepts

The concept ‘consciousness’:

Can we draw a boundary between animals which do and those that don’t have it? Are humans conscious when they are asleep and dreaming? What about sleepwalkers? At what stage does a foetus become conscious? Is it a matter of degree or are there discontinuities?

The concept ‘emotion’:

Confusions are shown by conflicting answers to these questions: Is surprise an emotion? If you love your country, is that an emotion, or an attitude? Can you have an emotion without being aware of it? Does an emotion have to have some externally observable/measurable physiological manifestation? Can a fly feel pain, or have emotions? Is there a stage at which a human foetus becomes able to have emotions? Could a disembodied mathematician have emotions?

General disagreement about criteria for adequately specified mental concepts:

Must there be externally detectable evidence (operationalism)? Must mental states be definable in terms of input-output relations? (behaviourism). Is introspection reliable evidence for existence of mental processes?

(Compare the general problem of how new indefinable theoretical terms are introduced in science, e.g. concepts referring to sub-atomic particles and forces.)

Attempt to get a broader view by exploring the space of architecture-based concepts

An instance of an architecture (+ the environment with which it interacts) generates a class of possible states, events, processes.

– An instance of a virtual machine architecture generates a class of possible VM states, processes, etc.

– Different architectures generate different classes of possible states, processes, etc. i.e. different VM ontologies.

– Observing actual phenomena in an architecture gives only a *partial* view of what the full range of possibilities for that architecture is. (Compare: the chemical compounds and chemical reactions found in nature are not the only possible ones.)

– Observing only *one type* of VM (e.g. only human minds) gives only a partial view of the full *range of classes* of possible states, processes, etc., in different organisms and machines.

– Leads to formation of concepts that cannot accommodate all important cases.

A Comparison: The Architecture of Matter

The current theory of the architecture of matter generates particular classes of concepts

- **of kinds of matter** e.g. atoms, molecules of various kinds, crystals, etc.
- **of kinds of events and processes** e.g. radioactive decay, chemical reactions.

- **Consider the way our concepts of kinds of matter, kinds of physical stuff, got extended and refined as we learnt about the architecture of matter.**
- **E.g. the periodic table of elements was explained by the theory of the architecture of sub-atomic physics.**
- **Understanding how atoms can and cannot combine generates a space of chemical concepts.**
- **There are concepts of types of *process* e.g. catalytic reaction, as well as concepts of types of *state*.**

Proposed generalisation: consider how a VM architecture generates classes of mental states, events, processes.

An architecture supports a class of concepts

As we get to understand the architecture of an operating system we find the need to introduce new concepts referring to states and processes that can arise through interactions between components of the architecture:

- Various notions of ‘load’ on the system
- The notion of ‘thrashing’
- The notion of ‘deadlock’
- The notion of ‘privilege’
- The notion of responsiveness

and more, especially as machines are networked.

Note that the fact that we can use numbers for some of these does not imply that the system has some kind of internal numerical variable representing those states.

(It may do if it does some self-monitoring!)

Extending Architecture-Based Ontologies

Some concepts are “generated” within an ontology that defines an architecture, e.g. for matter, for mental mechanisms, for social systems, for political systems, for a computer operating system.

- Science extends, corrects, and refines our theories of the underlying architectures.
- It thereby extends, corrects and refines our naive ontologies.
  - Some parts of an ontology may be discarded (e.g. ghosts, angels)
  - Other parts may be further subdivided and made more precise (e.g. carbon12, carbon14)
  - New parts may be added to the ontology (e.g. new elements, new states of matter, new phase transitions).
    - **Shallow extensions:** new concepts definable in terms of old ones.
    - **Deep extensions:** addition of new indefinable primitives.

NOTE: We constantly use ontologies referring to virtual machines with complex architectures, even if we are unaware of doing so (e.g. social, legal, political virtual machines – All ultimately implemented in physical machines. This leads to problems of supervenience, etc.)

We call them “machines” because they have components that interact.

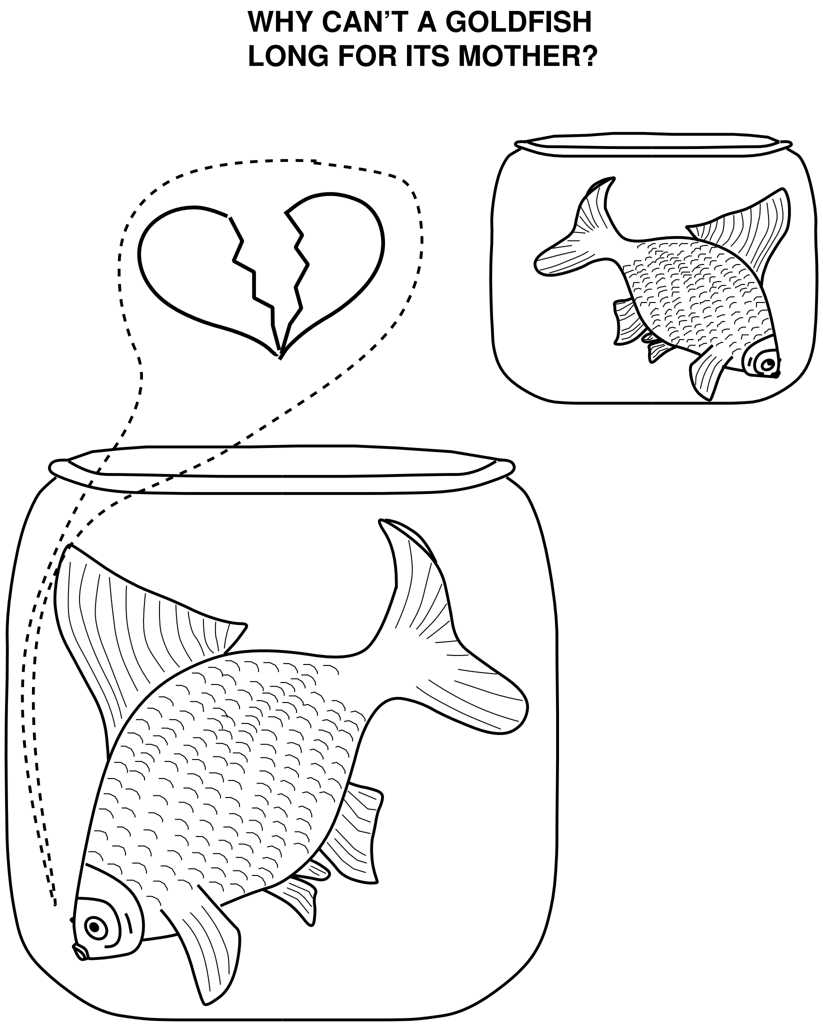
There are MANY kinds of information processing architectures – hence MANY sets of concepts

Physics studies ONE world with ONE architecture (several levels).

AI & cognitive science studies MANY (virtual machine) architectures for natural and artificial minds – supporting different sets of concepts of internal states:

- Flea minds (flea perception, motivation, emotions, consciousness)
- Mouse minds (mouse perception, motivation, emotions, consciousness)
- Cat minds ...
- Chimp minds ...
- Human neonate minds ...
- Your mind ...
- Minds damaged by Altzheimers' disease ...
- Long dead evolutionary precursors
- Various possible kinds of minds for robots and synthetic software agents.

Why can’t a goldfish long for its mother?



- Because it cannot make its mouth droop?
- Because it lacks tear glands to make it weep?
- Because it cannot sigh....?
- Because it lacks our proprioceptive feedback...??

No, because:

1. it lacks the appropriate information processing architecture
2. including representational mechanisms, concepts and knowledge.

Some of the requirements are discussed further in

http://www.cs.bham.ac.uk/research/cogaff/talks/#cafe04

Why call them all “minds”?

Why call different information processing systems “minds”? Does a flea have a mind?

Likewise is there any justification for using the same word “believe”, “emotion”, “learning” in talking about different architectures.

- These are terminological decisions and ultimately can be evaluated only on the basis of long term usefulness, e.g. in promoting development of explanatory theories, and perhaps reducing philosophical puzzlement.
- Whatever else human minds do they certainly process information: we acquire, store, interpret, transform, derive, search for, recombine, and use information of many kinds.
- We propose that it is useful to think of a mind as an information-based control system. (Sloman 1993).
- There are many sorts of information-based control systems (IBCS) some very simple, such as thermostats, others more complex including many sorts of organisms.
- Instead of arguing about which of these are and which are not minds, we can think of them as being more or less “mind-like”, and investigate their properties.

Using mental concepts for different sorts of minds

- Whether the same label (e.g. "desire", "belief", "intention", "emotion", "learning", "perception",) should be used for mechanisms, states, events, etc. in different VM architectures depends on whether it is useful to refer to common structural and functional features.
- There may be low level differences (e.g. in what is desired, and how desires are generated) along with high level similarities (e.g. in the way desires interact with percepts and beliefs to produce decisions, and behaviour).
- Sometimes there are both similarities and differences, e.g. between desires in an infant and desires in an adult human, or a mouse.
- When differences between features are important different words can remove ambiguity (e.g. believe$_{flea}$, believe$_{monkey}$).

There are important structural and functional similarities across different VM architectures, which the CogAff Schema, described later, helps to characterise. These may justify common labels.

We understand only a tiny subset of the space of possible virtual machine architectures

Different VM architectures are required for minds of different sorts

Compare

– **Minds for robots with bodies**
– **Minds for software agents**

We need to place the study of (normal, adult) human mental architectures in the broader context of

THE SPACE OF *possible* MINDS

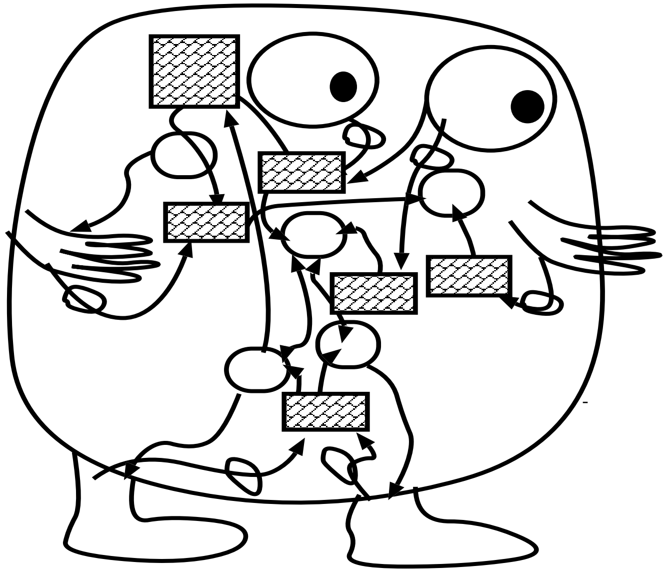
I.e. minds with different architectures that meet different sets of requirements, or fit different niches. Reject the standard philosophical assumption that there must be necessary and sufficient conditions for something to be or have a mind.

- **NB The vast majority of organisms do NOT have human-like architectures.**
- **Most have only more or less sophisticated (not stateless) reactive architectures. (e.g. single-celled organisms, insects, fishes, etc.)**

Residual problem: “Cluster concepts”

- Even if we specify relevant (VM) architectures and investigate the kinds of processes that can occur, that will not automatically define a *unique* set of concepts for describing the objects, properties, relationships, events, processes and states that can occur within the architecture.
- Working out a good way to organise a complex variety of phenomena can take time, and a lot of trial and error learning.
- The concepts in use at any time will often have many elements of arbitrariness and indeterminacy, because the need has not yet arisen to draw boundaries in various regions of conceptual space. Most of our concepts of naturally occurring phenomena are ‘cluster concepts’ referring to entities which can have various combinations of features and capabilities from a large set.
  - Objects with certain combinations of those features will be regarded as definitely instances, and objects with other combinations will be regarded as definitely non-instances.
  - But for many other combinations of features and capabilities there may be as yet no answer. E.g. “Is a flea conscious?” may have no answer.
  - But we can define precise new concepts: fleas have consciousness$_{XY}$, pain$_{PQR}$

What is the human architecture? Could it be an unintelligible mess?



YES, IN PRINCIPLE

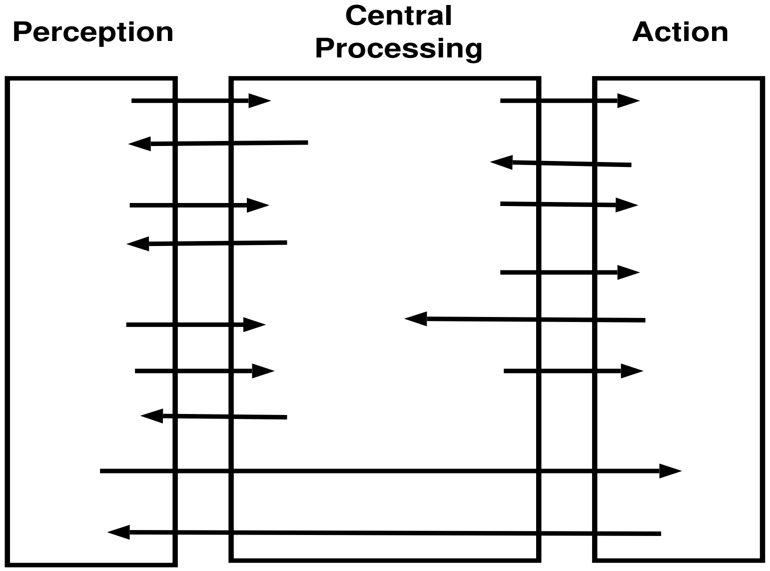
BUT: it can be argued that evolution could not have produced a totally non-modular yet highly functional brain.

Problem 1: The biological usefulness of *duplicate then differentiate*.

Problem 2: time required and variety of contexts required for a suitably general design to evolve. Re-combinable modules may help.

Problem 3: storage space required to encode all possibly relevant behaviours if there's no "run-time synthesis" module.

Towards A Unifying Theory of Architectures (For natural and artificial agents)



1. The "triple tower" perspective

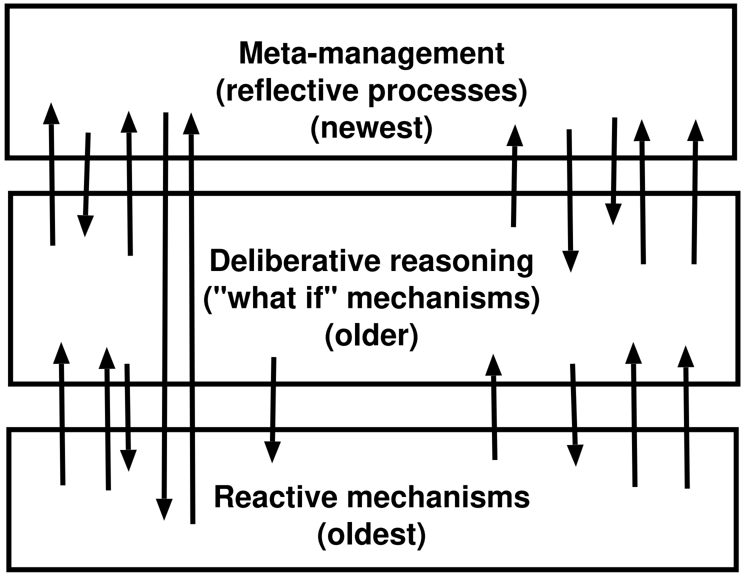
– perception,
– central states and processes,
– action mechanisms

(All with fuzzy boundaries)

(many variants —- e.g. Albus (1981), Nilsson (1998))

NOTE: we are not assuming that there is information-flow in one direction through the system. Many kinds of feedback and feedforward may be used in controlling the total system.

Another Common Architectural Partition



2. THE "TRIPLE LAYER" PERSPECTIVE – (FUNCTIONAL, EVOLUTIONARY)

The layers differ in:

– Evolutionary age (reactive oldest).
– Level of abstraction of processing (reactive least abstract),
– The types of control functions, and mechanisms used
  (e.g. ability to search, evaluate, compare)
– The forms of representation used
  (e.g. flat vs hierarchical compositional syntax and semantics)

(Many variants – for each layer)

Properties of the different layers

- **Reactive mechanisms** can be highly parallel, very fast, and use analog or digital components or both. Some **reflexes** (innate or learnt) connect sensors to motors.
  NB: Some reactions change only internal state.
- **Deliberative mechanisms** can represent and reason about **non-existent** or **future** possible actions or objects. Some can also reason about **what might have been** the case in the past.
  - Simple deliberative mechanisms may use only one step lookahead, and very simple selection mechanisms.
  - More sophisticated versions use compositional semantics in an internal language. They are inherently slow, serial, knowledge-based, resource limited. (Why?)
- **Meta-management** mechanisms can monitor, categorise, evaluate, and (to some extent) control other internal processes. They can also vary in sophistication.
  - **The evolution of sensory qualia:** occurs when it is useful for meta-management to look inside intermediate levels of perceptual processing (why?).

The CogAff schema combines the views: layers + pillars = grid

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

An architectural "schema" not an architecture: defines only *possible* components and links – not all need be in all organisms, or all robots.

Different architectures will have different subdivisions in all the boxes, and different connections between the components.

An 'ecosystem' of mind: A grid of co-evolving sub-organisms (cooperating and competing for resources), each contributing to the niches of the others.

The CogAff schema allows variants

Defines a set of types

- of component mechanisms
- of information linkages
- of control relationships
- of forms of representation

Different combinations will be present in different instances.

CogAff does NOT specify control flow, or dominance of control.

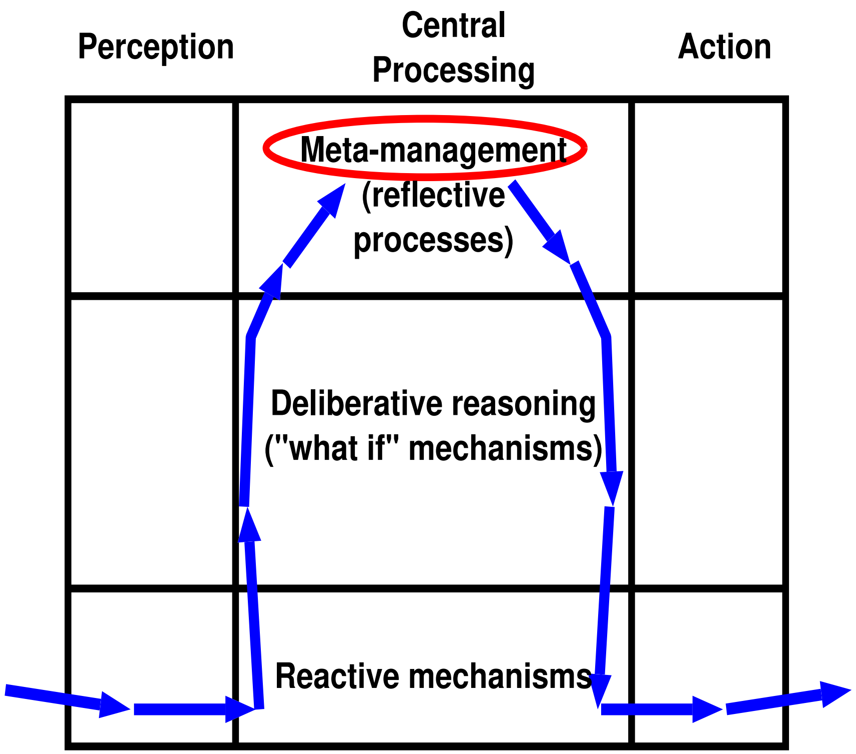
Many options are left open, including the possibility that many components operate concurrently and interact asynchronously sometimes overriding or ignoring information or instructions from other components.

The H-Cogaff architecture described later is a particularly rich special case, postulated as a model of human information processing.

But much simpler variants can explain other organisms and machines.

Each variant will define a particular family of types of mental states, processes, etc.

The popular ‘Omega’ (Ω) model

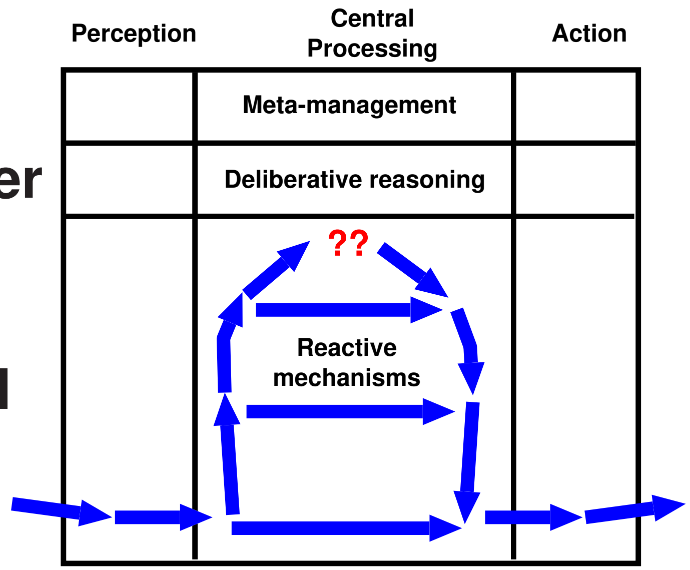


A simple pipelined information flow, with changing levels of abstraction – no layered concurrent perceptual and action towers separate from central tower. Much of CogAff omitted.

There are many variants, e.g. the “contention scheduling” model. (Cooper & Shallice, 2000)

Some authors propose a “will” at the top of the omega. (Albus 1981)

Another variant: Subsumption architectures



These use a hierarchy of reactive layers, which operate concurrently with lower levels subject to being controlled by higher level layers (Brooks 1991).

There is no deliberative mechanism, no form of representation with compositional semantics, only a limited sort of reactive meta-management.

Some champions of reactive subsumption architectures deny that humans use deliberative mechanisms employing internal representations as the environment is supposed to have all the information required. *How do they get to conferences, design computer programs, do algebra?*

Subsumption, like many other architectures, uses only a subset of the mechanisms allowed in the CogAff schema. We should avoid all dogmatism and ideology, and investigate which subsets are useful for which organisms or machines, and how they might have evolved.

CogAff allows layered perception and action

- In many models perception and action systems are mere transducers (e.g. in the Omega model):
  'Peephole' models of perception and action.
- The CogAff schema allows perceptual and action systems to have multiple layers operating concurrently, with information flowing *directly* to and from different layers in the central pillar:
  'Multi-window' models of perception and action.
- Examples:
  - Posture control and fine-grained control of feedback require very low-level visual information (e.g. measures of optical flow, angular distance) to be fed to reactive mechanisms.
  - Concurrently, information about hierarchically organised spatial structures (furniture, doorways, possible routes) may be sent to path-planning deliberative mechanisms.
  - Reading hand-written text, hearing speech and sight-reading music require a mixture of bottom up and top down perceptual processing involving several layers of abstraction. So the perceptual architecture must be in part produced by learning.

Conjecture: Perception of mental states

- In humans (and some other animals?) perceptual mechanisms evolved to use rich mental ontologies

  If you are likely to be eaten by X what is more important for you to perceive

  – The shape of X's body?
  – Whether X can see you?

- How? If meta-management mechanisms used a mental ontology for self-categorisation, the same ontology might be linked to perceptual mechanisms if other agents give sufficient evidence of mental processes in their behaviour.

  – Some of this is inherent in the behaviour of intelligent organisms: the direction of gaze, incipient movements etc.

  – Some may be a product of co-evolved involuntary expression and perceptual algorithms (e.g. smiling, frowning).

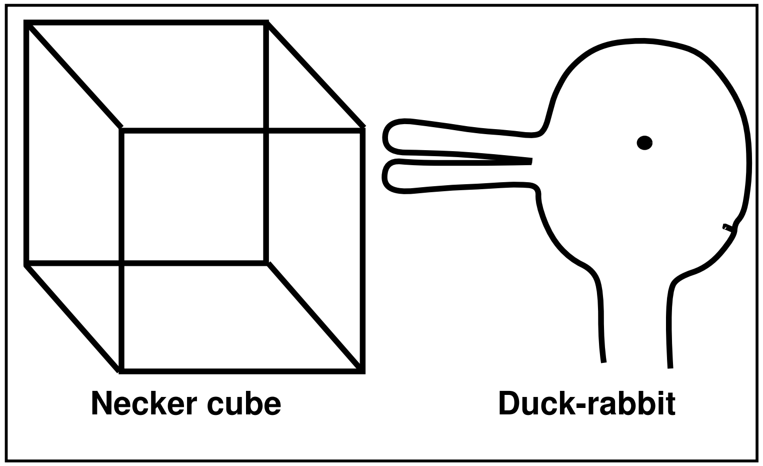
- I.e. Primitive implicit theories of mind probably evolved long before anyone was able to talk about theories of mind.

  So evolution solved the "other minds" problem before there were any philosophers to notice the problem.

The main functions of vision

Marr (1982): to inform the organism about shape, motion and colour: i.e. *geometrical and physical properties*.

Gibson (1979): to provide information about (positive and negative) *affordances*: support, graspability, obstruction, passage, etc.



Evidence for different processing levels in perception can be found in the variety of visual ambiguities.

Seeing the switching Necker cube requires geometrical percepts.

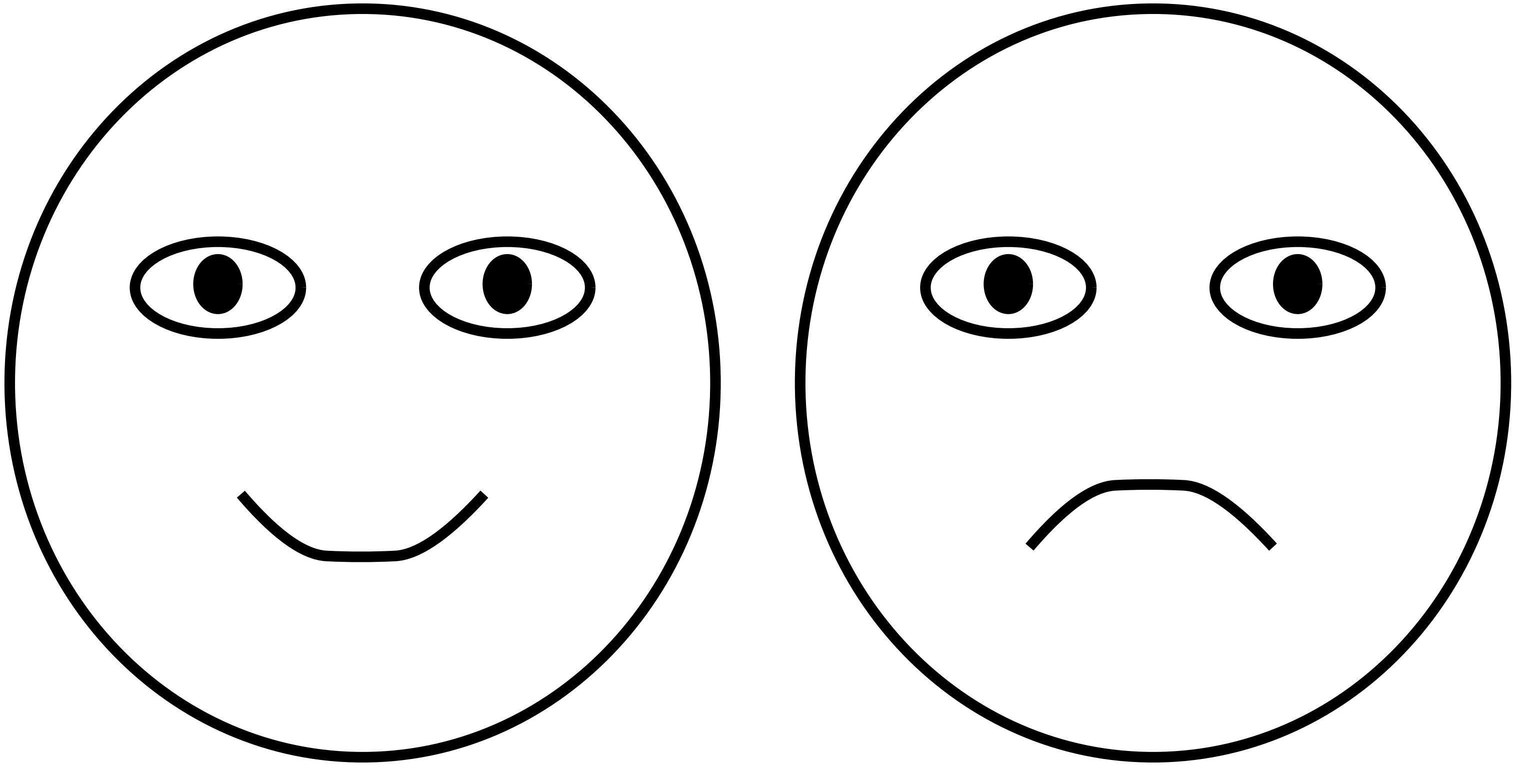
Seeing the flipping duck-rabbit uses far more subtle and abstract percepts, going beyond geometric and physical properties.

Things we can see besides geometrical properties:

– Which parts are ears, eyes, mouth, bill, etc.
– Which way something is facing
– Whether someone is happy, sad, angry, etc.
– Whether a painting is in the style of Picasso...

Perception of faces

What do you see? Is it purely geometric?



ARE THE EYES THE SAME OR ARE THEY DIFFERENT?

Perception: more than describing what's "objectively out there"

Concurrently active internal processes may have diverse needs, which must be served in parallel by perceptual mechanisms, or through sequential switching of resources.

- Posture control, face recognition, route planning, can all be served in parallel by visual mechanisms.
- More generally, perceptual needs of a subsystem are defined not by the physical/geometrical nature of the environment, but by the functions of the subsystem and its capabilities, including its processing capabilities.
- Therefore "affordances" available to an animal are a function of the sub-system that uses them, not just features of the environment.

  Different sub-systems use different affordances, and different formalisms and ontologies. *(Evidence from brain damage.)*

  Think of a mind as a virtual machine architecture containing an "ecosystem" with many co-evolved interacting sub-species.

(Computation may not be the best model for this.)

Different perceptual needs define different affordances

Examples:

- Online or ballistic control of external actions or triggering of physiological (e.g. sexual) responses in a reactive subsystem
- Consideration of possible actions in a deliberative subsystem.
- Triggering of sympathy, etc. in social animals
- evaluating what is detected,
- triggering new motivations
- triggering “alarm” mechanisms (described below)
- . . . . .

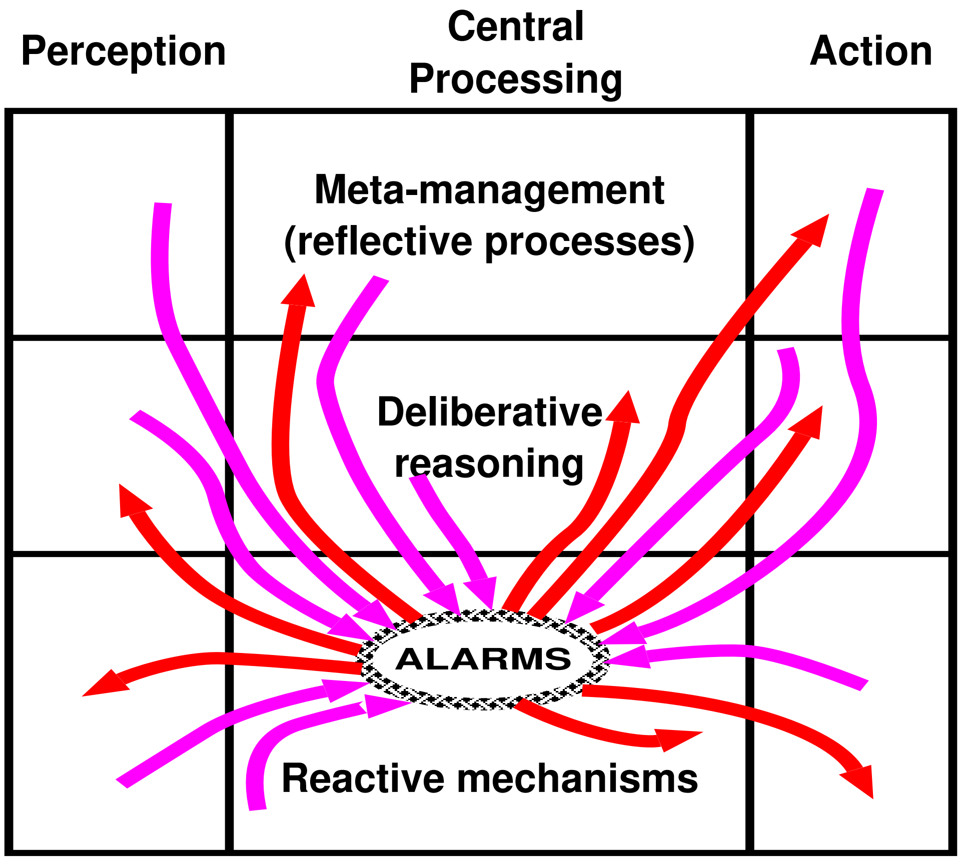
These all need internal languages of some sort to represent the information acquired. The type of language will differ for different subsystems in the same agent.

Some perceptual systems may be linked to “alarm” mechanisms (described later).

As processing grows more sophisticated, so it can become slower, to the point of danger.

REMEDY:

FAST, POWERFUL, "(RELATIVELY) GLOBAL ALARM SYSTEMS"



Not all global: some specialised (e.g. blinking reflex)

Features of alarm systems

Alarm mechanisms must use fast pattern-recognition and will therefore inevitably be stupid, and capable of error!

NB: An alarm mechanism is just part of the reactive 'layer'. Drawing it separately merely helps to indicate a special role for a subset of reactive mechanisms.

Many variants of alarm systems possible. E.g.

- **Purely innate or trainable (e.g. athletes, surgeons).**
- **One global alarm system or several "less global" specialised alarm subsystems:**
  - **Different specialised 'alarms' concerned with the reactive mechanisms for body control, e.g. posture, withdrawal from hot or sharp objects.**
  - **Specialised 'alarm mechanisms' monitoring processes within the deliberative system.**
  - **Alarm mechanisms may differ in what they can control. Compare posture adjustments *vs* blinking reflex *vs* causing a saccade, *vs* causing thoughts to be re-directed (control of meta-management by alarm mechanisms)**

(Spinal reflexes, brain stem, limbic system, ...???)

Architecture-based concepts: Alarms and emotions

Conjecture: Many of the things we call emotions are a result of the operation of something like alarm mechanisms, triggered by detection of external or internal states or events requiring (more or less) global redirection of control.

– **Primary emotions:** originally exclusively involved mechanisms in the reactive layer. But in animals with other layers, they can have additional side-effects.

– **Secondary emotions:** triggered by events in the deliberative layer (realising what might happen, or what might have happened)

– **Tertiary emotions:** events triggered in various parts of the system interfering with meta-management, e.g. causing loss of control of attention (short term: embarrassment, long term: grief).

Many more cases can be distinguished, by analysing possible perturbations in greater detail.

Compare: Simon 1967, and the Cogaff project directory.

NOTE: partly similar roles within the global architecture justify similar labels for processes (e.g. "primary emotion", "hunger") in organisms or machines with different architectures.

Extra components for human-like systems

The 3 x 3 grid (with alarms) gives only a very crude categorisation of components of an architecture. Many further sub-divisions are required to meet the specific needs of different organisms, robots, etc.

In particular, deliberative and meta-management mechanisms can vary enormously in their sophistication.

Extra mechanisms could include:

- **Personae:** Different modes of operation for meta-management, etc.
- **Standards, values:** For high level control of some decision making.
- **Categories:** Used by perception, deliberation, self-monitoring,learning planning.
- **Motives:** Current short term and long term springs of action.
- **Attitudes:** Enduring semantically rich clusters of factual and control information.
- **Formalisms:** For expressing percepts, goals, knowledge, plans, etc.
- **Skill-compilers:** Allowing deliberative processes to “train” reactive ones.
- **Arbitration mechanism:** for resolving conflicts
- **Filters:** E.g. dynamic thresholds can protect resource-limited deliberations.
- **Moods:** Context-sensitive mechanisms for global state (various durations).

Evaluators of various kinds

- Current state can be evaluated as good, or bad, to be preserved or terminated.
- These evaluations can occur at different levels in the system, and in different subsystems.
- They have different control functions, e.g.
  - Immediate control: termination, preservation, initiation, modification of a state of affairs or current behaviour.
  - Long term control: learning by changing likelihood of future responses
  - Spatially directed responses (e.g. removing hand from heat)
  - Undirected behavioural responses
  - Triggering debugging of a strategy
  - Triggering behaviour towards others: punishment, submission, comforting, etc.
- A rich enough architecture can account for many different kinds of pleasures and pains.

  (These are often confused with emotions.)

Varieties of motivational sub-mechanisms

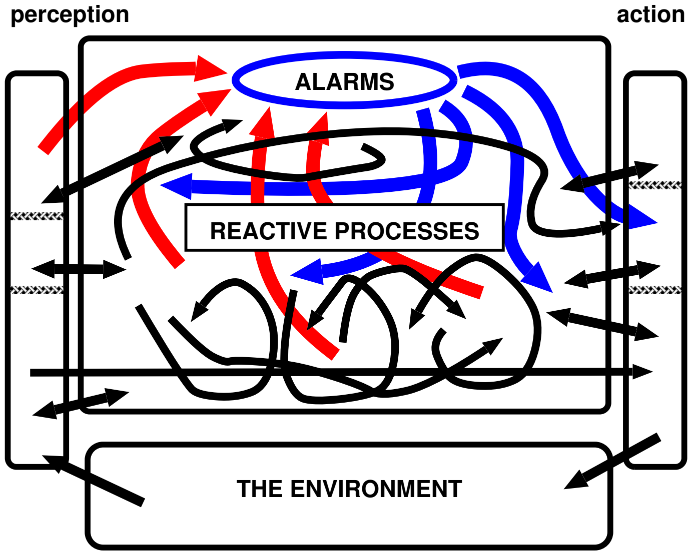
Motivation comes in many forms (depending on the needs and capabilities of the architecture):

- Motives, goals and related items, e.g. preferences, can be short term, long term, permanent.
- They can be triggered by physiology, by percepts, by deliberative processes, by meta-management.
- They can be part of the reactive system, part of the deliberative system, part of meta-management.
- They can be explicit or implicit in how things work.
- They can be syntactically and semantically simple or arbitrarily complex (e.g. the full specification for the house of your dreams).
- They can play roles of varying complexity in the generation of internal and external behaviours, e.g. triggering processes like
  - considering whether to accept/adopt the goal
  - considering when or how to achieve it
  - considering when to consider it!
  - evaluating a plan
  - evaluating execution of a plan

Motive generators

- There are many sorts of motive generators: MG
- However, motives may be in conflict, so Motive Comparators are needed: MC
- But over time new instances of both may be required, as individuals learn, and become more sophisticated:
  - Motive generator generators: MGG
  - Motive comparator generators: MCG
  - Motive generator comparators: MGC
  - and maybe more:
    MGGG, MGGC, MCGG, MCGC, MGCG, MGCC, etc ?
- What are the architectural requirements for support of this kind of richness of motivation?
- How much of it do humans and other animals actually have?
- How much of it will intelligent robots and software agents need?

Not all parts of the grid are present in all animals

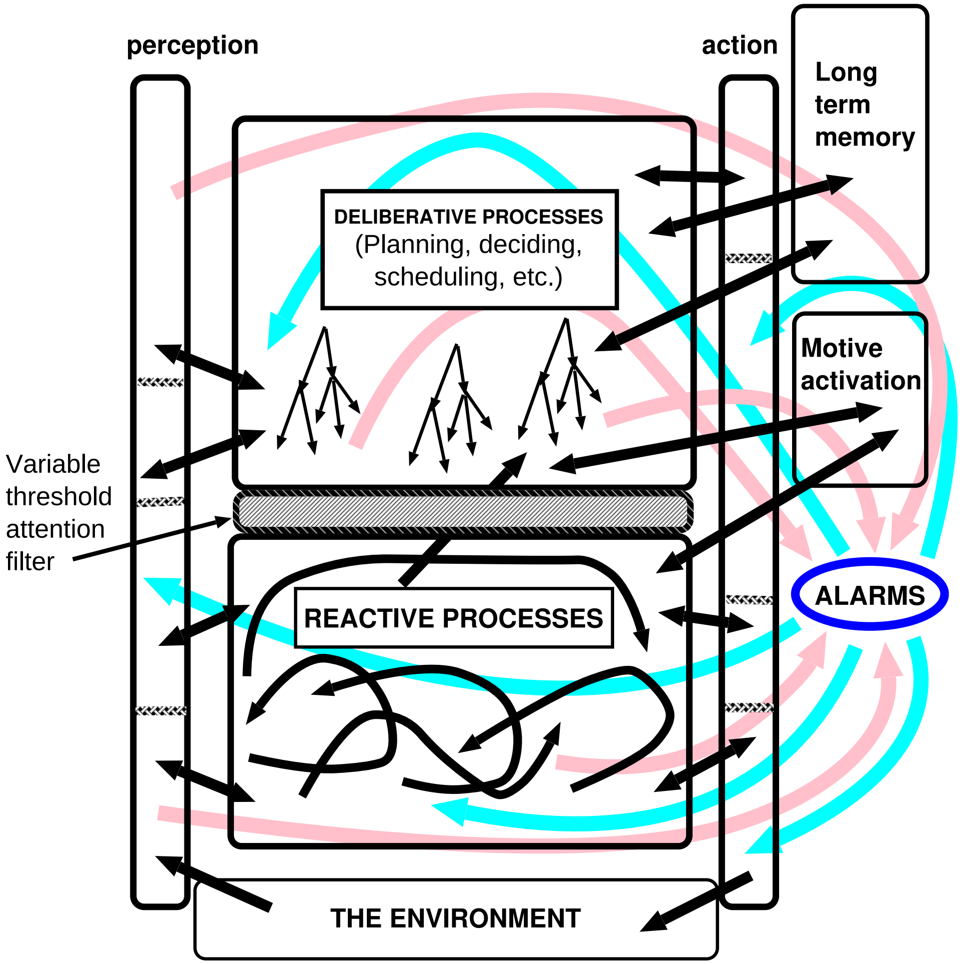


HOW TO DESIGN AN INSECT?

Individual insects are reactive (i.e. they lack deliberative mechanisms), though their reactions can change internal states. However, group behaviour can give the appearance of exploration and deliberation, e.g. in the construction of termite 'cathedrals'.
(A primitive deliberative VM?)

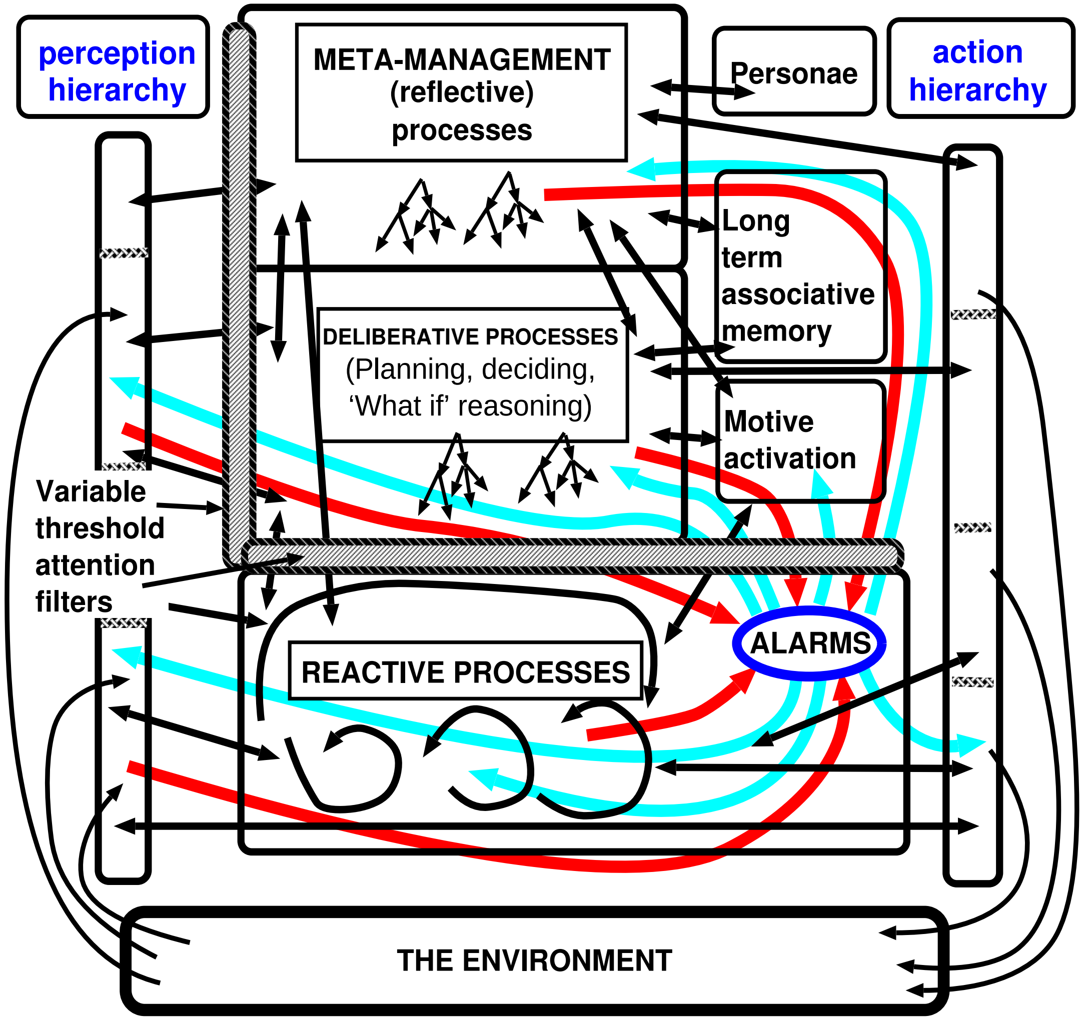
But there is no evidence that any individual insects understand what is being constructed or how.

Add a deliberative layer, e.g. for a monkey?



Adding full deliberative capabilities requires substantial changes in the architecture and possibly considerable cost, e.g. in energy.
Only a tiny subset of species have found this worth while!

H-COGAFF: A human-like architecture



A hypothesised human-like instance of the CogAff schema, described in more detail in Cogaff papers: http://www.cs.bham.ac.uk/research/cogaff/

Many profound implications

Comparing VM architectures we find many profound differences in the structures and processes they support e.g. different kinds of:

- **perceptual processes**
- **goals that can be generated**
- **action proposing mechanisms**
- **action selection**
- **arbitration (conflict resolution)**
- **possible varieties of learning and development**
- **moods, emotions, and other affective states**
- **Possible effects of brain damage,**

  The more complex the architecture, e.g. the more concurrent independently active components there are, the more subtle and varied the effects of different sorts of brain damage. E.g. some architectures allow damage of one component to be partly compensated for by changes in others. Some allow this only after new learning. Damaging one visual subsystem may leave another one functioning normally, as in 'blindsight'.

THERE IS MUCH WORK TO BE DONE

There are many parts of the system about which little is known.

- In particular we conjecture that one of the biggest gaps concerns vision, and more generally the grasp of spatial structure.
- Humans and many other animals e.g. nest-building birds, berry-pickers, tree-climbers, tool users, hunters of various kinds, can apparently to take in and use accurate information about intricate spatial structures at very high speeds, e.g. when birds fly through branches or squirrels or monkeys travel through tree-tops.
- In humans (and perhaps some other animals) this grasp of space and motion extends to being able to visualise complex structures and motions in solving problems, e.g. working out how some mechanism operates, and how to fix it, and much mathematical reasoning.
- This seems to be related to our ability not only to see *what exists* but also *which changes are possible*, which is crucial to understanding our environment and how to act in it.(Sloman 1996b)
- We also don't know how much of the architecture uses analog (continuous) as opposed to discrete information processing, and how much influence that has on the whole system.

We also don't know how to put all the pieces together.

PART FOUR

MORE ON CAUSATION, VIRTUAL MACHINES AND LEVELS

- Causation and counterfactuals
- Varieties of emergence
- Physical determinism does not imply backward causal closure
- Our normal concept of causal allows overdetermination of effects.
- This explains how both virtual machine phenomena and the physical substrate can be causally efficacious.
- How can virtual machine states have semantic content?
- Mind and computation (two notions of computation)
- The ‘design stance’, directed to virtual machine architectures, makes the ‘intentional stance’ unnecessary.
- How machines will have qualia

Ontology, Causation, Virtual machines 1

- We need to study ontologies that involve causal interactions between components of a virtual machine. E.g.
  - A parser in a compiler interacts with error handler and code generator.
  - The process scheduler and memory manager in an operating system interact with each other.
- This presupposes a notion of “causation”. Analysing that concept is one of the hardest unsolved problems in philosophy.

Ontology, Causation, Virtual machines 2

This presupposes a notion of "causation".

Analysing that concept is one of the hardest unsolved problems in philosophy.

- **For a subset of theoretical computer science, causation is irrelevant: a computation is just a mathematical structure (possibly infinite), something like a proof (often a computation is seen as a type of proof.) It need not occur in time, or have any causes or effects. (E.g. a Gödel number.)**

  E.g. notions of space complexity and time complexity refer to purely syntactic properties of a "trace" of a program execution: which can be viewed as just a mathematical structure.

- **For software engineers, robot designers, and computer users, computation involves a process in which things exist, and events and processes happen.**

A causal paradox

- We tend to assume that physics is causally closed backwards

  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 can intervene.
- We assume that events in virtual machines can cause other events in the virtual machines, and can also produce physical effects
- So events in virtual machines can cause physical events.
- So physics is not causally closed after all??
- Or perhaps our desires do not cause our actions??

Some people have assumed that we need some causal gaps (such as quantum mechanical indeterminacy) in physics to enable mental phenomena or other virtual machine phenomena to have causal powers.

I.e. if all causes of physical events are physical (backwards causal closure of physics) then minds and computational virtual machines must either be physical after all (the mind-brain identity theory), or else epiphenomenal. Where's the fallacy?

Can Virtual Machine events be 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.

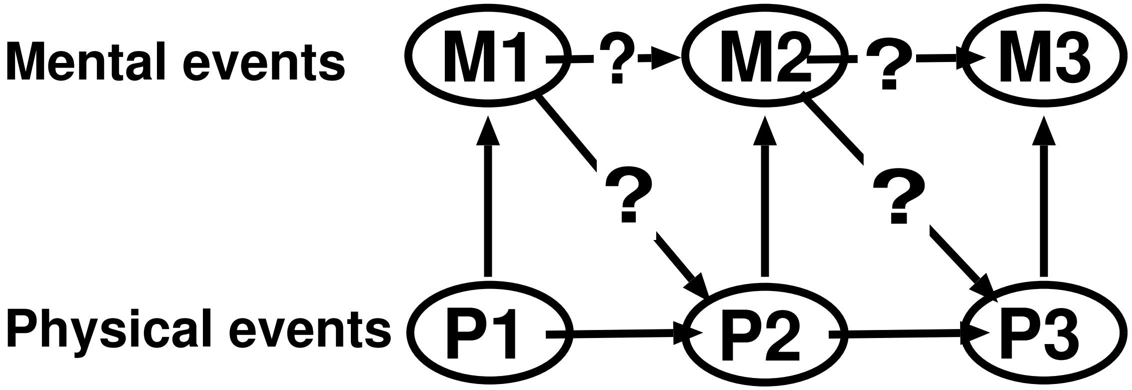
How can that be, if all these non-physical phenomena are *fully implemented* in physical phenomena?

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.

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

This would rule out all the causal relationships represented by arrows with question marks.

Answer: Non-physical causes are possible

Problems with the 'monistic', 'reductionist', physicalist view that non-physical events are epiphenomenal:

- **It presupposes a layered view of reality with a well-defined ontological bottom level. IS THERE ANY SUCH BOTTOM LEVEL?**
- **There are deep unsolved problems about which level is supposed to be the real physical level, or whether several are.**
- **It renders inaccurate or misleading much of our indispensable ordinary and scientific discourse, e.g.**
  - Was it the government's policies that caused the depression or would it have happened no matter which party was in power?
  - Your anger made me frightened.
  - Changes in a biological niche can cause changes in the spread of genes in a species.

Some philosophers try to reconcile this by adopting mind-brain identity, or Virtual Machine/Physical machine identity thesis.

But identity is hard to square with apparent asymmetry.

E.g. are physical events implemented in mental ones?

The identity theory trap

Some philosophers attempt to remove the puzzle by saying that M1 and P1 are the same thing: i.e. adopting the mind-brain identity theory, or VM-computer identity theory.

- **This can cause problems if you want to be able to talk about identity of virtual machines across alternative possible physical states (e.g. "What would the operating system have done about allocating memory if process P25 had terminated just before P32 requested additional memory, instead of after?")**
- **It is not clear that discussions regarding identity are discussions of substance: what is treated as identical with what may be partly a matter of convenience, or conceptual clarity, rather than truth.**

  E.g. the ancient Greeks noticed that it is not clear what is meant by saying that you do or do not step into the same river twice.

  If you borrow my axe, then replace the blade because it breaks, and later replace the handle because it breaks. Is what you return the axe I lent you? How much must be the same?
- **Solving any philosophical problem by saying A is *identical* to B, risks vacuity because of the indeterminacy of the notion of identity.**

Conjecture: Towards a schema for causation

In the "everyday" ontology, used in our practical interactions with one another we use a notion of "causation" that is POLYMORPHIC.

"X caused Y" does not have a fixed, context-independent meaning.

There is a schema, which has to be filled out differently in different contexts, even if Y is the same.

"Redundant causation" is the norm.

Redundant (multiple) causation: examples

- We can correctly say of a particular person that his death was caused by smoking, that his death was caused by lung cancer, or that his death was caused by certain physiological processes that occurred in the last few minutes of his life. The assertions do not contradict one another. (Why not?)
- We can say that a certain car crash was caused by poor driving or by ice on the road, depending on who is asking and why.

These statements, though both true, are relevant to different contexts of enquiry. E.g

- Asking why the driver crashed this time but not when he drove on this road previously could be answered by saying the crash was caused by ice.
- Asking why this driver crashed but other drivers on that road did not crash could be answered by saying that this person was a poor driver.

Redundant causation is the norm.

Each question about causation is linked to a range of possible circumstances (same driver, different occasions, different drivers same physical conditions, etc.).

Even if Y is a particular event, different answers to "what caused Y" are relevant in different question-asking contexts.

THERE IS NO UNIQUE, GLOBAL, CONTEXT DETERMINING WHICH STATEMENTS ABOUT CAUSAL CONNECTIONS ARE TRUE.

Compare: there is no unique global context determining which things are better than others.

I.e. both

> **"X CAUSED Y"**

and

> **"A IS BETTER THAN B"**

implicitly refer to some further context which determines the conditions under which the statement is true or false.
(E.g. Better for what? In what respect?)

Causation is a relation of more than two items

So there is no uniquely correct, context-independent, answer to the question: "Did X cause Y?"

There is some implicit or explicit context, usually involving a practical or ethical or legal question, which determines which factors are relevant to answering the question.

So "X caused Y" is not just a statement about events X and Y. *There is implicit reference to some context.*

- **If Y is Fred's car crashing, whether X should be a collection of physical features of the car and the road, or Fred's laziness about maintenance, depends on who is asking the question, and why.**
- **When the context is unspecified, disputes about causation can be at cross-purposes, lacking any correct answer.**
- **Context is important because the question is not a purely factual one, but is relevant to practical decision making.**

NOTE: the concept "cause" is not an invention of physicists.
Insofar as biological organisms are built to learn and use information about causal relationships, the concept is a product of biological evolution, not physics (or philosophy).

Conjecture: What “X caused Y” means.

The meaning of “X caused Y” is quite complex, with many unobvious and subtle features. We conjecture that it has three main components of the form:

1. **X happened and Y happened.**
2. **In a certain variety of possible circumstances, C1, if X had happened then Y would also have happened.**
3. **In a certain variety of possible circumstances, C2, if X had not happened and nothing else had occurred capable of producing Y, then Y would not have happened.**

CHANGE THE TENSE FOR ONGOING OR FUTURE CAUSES.

Which sets of circumstances C1 and C2 are relevant will depend, in subtle and complex ways, on the practical context in which the question about causation is asked.

E.g. attempting to assign blame leads to different questions from attempting to decide how to behave in future.

Even if no explicit question has been asked, statements about causation presuppose a type of question that they are answering.

(In this they are like “A is better than B”.)

Defeasibility of statements about causation

Any attempt to specify precisely the circumstances relevant to "X caused/causes/will cause Y" can be rebutted with a refined specification which makes the consequent "Y would have happened/is happening/will happen" false.

- You may have good reason to think that in circumstance C1, if X had happened then Y would have happened. E.g. if Fred had drunk less he would have avoided the crash.
- But you may not have considered what could occur if Fred had a heart attack, or if aliens from another planet with very advanced technology had intervened.
- A disputant may or may not be able to persuade you that a previously unnoticed possible situation is relevant: depending on your high-level practical goals. E.g. trying to prevent disasters in the next 20 years is not the same as trying to prevent disasters in the next 2000 years.
- Statements about causation, like statements about counterfactual conditionals, are inherently (partly) indeterminate in meaning.
- In general it is impossible to produce a non-trivial, non-circular, context-independent, specification of the relevant variety of circumstances.

NOTE: Specifying the circumstances as those in which X suffices for Y is, of course, circular if you are arguing about whether X caused Y.

Multiple realizability and causation

When asking if VM event X caused Y, the difficulty in specifying the classes of circumstances to fill in schemata 2 and 3 is compounded by multiple realizability of virtual machine states and processes.

We may be unable to specify the physical circumstances in which X will occur, let alone those where X will produce Y.

If X is a chimp's decision to select one berry rather than another, there are myriad circumstances in which that decision would be followed by the action of picking up the berry, because organisms have many interacting mechanisms (including perceptual and motor control mechanisms) produced by evolution specifically to *ensure* that decisions are carried out, if necessary by counteracting or compensating for many possible perturbations during the process.

But (apart from relatively simple homeostatic mechanisms) we usually don't know precisely

– what the mechanisms are, or
– what the variety of circumstances is in which they suffice for their biological function, nor
– how various kinds of growth, learning, or damage-repair will modify the underlying physical implementation, nor
– how the implementation can vary from one member of the species to another.

Similar problems arise for VMs in computers

Suppose we know that an event in a VM in a computer (e.g. an attempt to access a file) will cause some other event (e.g. checking the access rights of the program).

We may not know or be able to predict all the current and future technologies that could produce a physical implementation of such processes, nor the variety of types of intrusions that could interfere with normal functioning of the mechanisms, nor their likely effects.

If the computing system is the result of design and implementation work done by different people solving different sub-tasks, or if the system has done some self-optimisation or self-modification (e.g. self-tuning schedulers or file managers), then our ignorance is partly like our ignorance about biological designs.

In both cases we don't know precisely which range of circumstances we are quantifying over.

Despite all that, I can be confident that the Lisp VM I am currently using includes a sorting mechanism so that if given the list [3 99 1 5 6] it will return [1 3 5 6 99].

Is such confidence justified? Usually, but occasionally misplaced.!
We don't know all contexts in which things can go wrong

Causation has some counterintuitive properties

"Causes" as (partially) analysed above is not in general a transitive relation, because different sets of circumstances can be referred to when we say that A causes B and B causes C.

Suppose a person X has a fall, producing a fractured bone.

Then it may be natural to say:

- X's fracture causes him pain
- X's being in pain causes Y to feel unhappy

but it can be at least misleading to say that

- X's fracture causes Y to feel unhappy,
  because there are too many ways in which the fracture might have occurred without Y being unhappy and too many ways in which Y might have been unhappy even if the fracture had not occurred.

Whether such a transitive inference from X caused Y and Y caused Z to X caused Z is valid may depend on the sorts of contexts in which the first two relations are considered. If the same sets of conditions are relevant to both, then the third relation holds.

Generally, the more intervening steps the more shaky the inference.

Multiple causes of the same event are possible.

- That has already been illustrated with the smoking and car crashing examples.
- E.g. we could say that the ice on the road and the poor driving caused different aspects of the crashing event.
- Over-determination often involves multiple aspects.
- Similar remarks apply to physical events (e.g. walking) which are caused both by mental events (e.g. deciding to leave the room) and physical events (e.g. previous states of the person's brain and the perceivable environment).

  NOTE:

  For a related, but different view, emphasising Bayesian probabilities, see Pearl 2000.
  Summary at Judea Pearl's website:
  http://bayes.cs.ucla.edu/jp_home.html

How P1 and M1 can both cause P2:

That M1 causes P2 (or some aspect of P2) is not refuted by P1 causing P2, because even if P1 did cause P2, it may still be true that:

1. M1 happened and P2 happened
2. In a certain variety of possible circumstances, C1, if M1 had happened then P2 would also have happened.
3. In a certain variety of possible circumstances, C2, if M1 had not happened and nothing else had occurred capable of producing P2, then P2 would not have happened.

2 and 3 are correct only because the variety of ways in which M1 can exist or not exist is constrained: the physical conditions under which M1 exists, and the variety of conditions in which M1 does not exist are limited.

M1 could be kept true, or made false only in ways that also keep P2 happening, or prevent it happening. Thus we get no contradiction between the above and these:

4. In a certain variety of possible circumstances, C3, if P1 had happened then P2 would also have happened.
5. In a certain variety of possible circumstances, C4, if P1 had not happened and nothing else had occurred capable of producing P2, then P2 would not have happened. I.e. both M1 and P1 can be causes of P2.

Example: Causation in control systems.

- In a computer controlled chemical plant, the machine takes a decision: an event in the virtual machine M1, causes a later physical event P2, such as a valve being opened.
- However, an earlier physical event P1, involved in the implementation of M1, is *also* a cause of P2.
- There is no contradiction here, given the normal interpretation of 'cause'. This sort of multiple causation is commonplace in the engineering world.
- Often the only relation of interest to the engineers is the relation between the VM events and the physical events, e.g. because the VM process involves a software bug which has to be removed, or because the VM can be generalised to deal with more situations.
- The precise physical details when the VM is running with the bug may vary and those when it runs after the bug has been fixed may vary. Software engineers often neither know nor care about them.
- However, they would care if a physical fault, e.g. a memory fault, causes the event P2 not to occur, or to occur in an undesirable modified form.

Our ignorance about brains

- Likewise, we rarely know or care about events in our brain, when things are normal.
- But we do care about brain events when there's damage or disease.
- This commonplace view of "biological mental causation" (in humans and animals) seems to parallel the case of "artificial mental causation" (i.e. causation in software virtual machines).
- At present the latter are simpler and easier to understand than the latter.
- So if we analyse carefully the products of engineers and scientists building working models and systems that control complex machinery, we may be able to develop a conceptual framework that enables us to ask, and perhaps answer, refined and clarified versions of old philosophical questions.
- It is also necessary to get clearer about counterfactual conditionals, and explain why the "politician's semantics" for counterfactuals is incorrect.

  I.e. when someone says "What would you do if XYZ happens?" the politician answers, inappropriately, "XYZ won't happen". (Not only politicians!)

Summary so far: VM events can be causes

If the arguments so far are correct then

- the assumption that backward causal closure in physics follows from causal determinism is false, because
- the same thing can have multiple sufficient causes at different levels, answering different questions
- therefore the arguments to prove the VM events cannot be causes, and must be epiphenomenal are unsuccessful.

We do not claim to have a knock-down proof

Our analysis of causation is still in need of further elaboration and critical assessment.

If our analysis is correct, then

- **Human minds' being essentially information processing virtual machines does not imply that they and their contents are epiphenomena, any more than poverty, crime, social inequality, war or the spread of rumours are epiphenomena.
  They are all real and they can all have effects.**
- **This does not require adoption of a mind/brain identity thesis (or a poverty/physics identity thesis).**

(Compare philosophical work by Peter Lipton at Cambridge, on causation and dispositions.)

Causation in complex virtual machines

There are still things to be explained about how virtual machines, including minds, relate to the machines in which they are implemented, including lower level VMs or physical machines.

A virtual machine architecture can include very large numbers of components that are

– concurrently active,
– constantly interacting,
– sometimes competing and sometimes collaborating
– with many sorts of short term and long term feedback loops
– some analog (continuous) and some digital (discrete)
– some synchronised and some asynchronous

So the collection of counterfactual conditional statements true of such a system will be very complex, and possibly very hard to discover.

Even systems we have designed may modify themselves (including reprogramming themselves) so much that we fail to understand how they work.

Computation with and without causation.

Returning to an earlier theme: What is computation?

- The theoretical/mathematical notion of computation refers to a purely formal class of structures. Whether anything in the universe does or does not have that structure is irrelevant.
- For a more common notion of computation, employed by *designers* and *users* of computers, the notion of causation is central: computers are machines that do things, externally or internally.
  - What is done internally, within the virtual machine, (e.g. reversing a list, updating a database, compiling a new subroutine, altering weights in a network, keeping scheduling statistics) need not produce external behaviour, though it can change the possibilities for future behaviour (internal or external), e.g. the questions that can be answered (though if not asked they will never be answered).
  - So the results of purely internal operations can change the causal powers of the machine. I.e. some machines have causal powers to change their causal powers.
  - This is true of many machines: e.g. temperature compensation mechanisms in a clock, and unwanted changes in accuracy or reliability that result from use of a machine (worn parts).

Machines with semantic content

We need to understand what it is for a machine to use X to refer to Y, i.e. to treat something as having *semantic content*.

- **Virtual machines often manipulate *information about something*. E.g. The machine may be controlling a chemical plant, or an aeroplane, or answering questions about company employees.**
- **Such machines are descendants of much older machines designed:**
  - **To control other physical machines,**
    e.g. looms used in weaving
  - **To perform operations on abstract entities,**
    e.g. arithmetical calculations on census data, or simply on numbers.
- **Jacquard looms, musical boxes, and later on Hollerith machines, showed how a machine could have a fixed part and a variable part containing 'instructions' (e.g. punched cards, or rotating discs).**
- **The development of electrical machines made it much easier to extend the class of operations on abstract entities to include operations which changed the machine's own future operations.**
- **However Babbage and Lovelace had that idea much earlier – long before Turing and von Neumann.**

Mind and computation

The key notion of mind is also the processing of information, in

– percepts,
– beliefs,
– desires,
– memories,
– skills,
– hopes,
– fears,

etc.

- Processes that involve acquiring, storing, transforming, interpreting information are not *physical*: they involve non-physical entities, e.g. numbers, words, rules, images, procedures, etc. However they are *implemented* in and dependent on physical machines.
- So the ‘Physical symbol system’ idea of Newell and Simon is confused because the important symbols are non-physical symbols in *virtual* machines.

  It should be: “Physically implemented symbol system”

Minds, causation and computation

We have drawn attention to the following

- Virtual machines and their components can have causal powers, including the ability to produce internal and external changes.
- They are fully implemented in physical systems, but the implementation relationship is very complex, may go through several levels of virtual machines, and instead of fitting simple notions of correspondence between VM and physical components may depend on subtle and complex mappings managed by software (e.g. interpreters, compilers, schedulers, memory managers, interrupt handlers, etc.)
- The space of possible VM architectures is huge and very diverse: we understand only a tiny fragment of it. We should keep an open mind about new varieties that will be discovered, e.g. including new mixtures of digital and analog, synchronous and asynchronous, local and global, processing of information and control.
- Some of the architectures we have begun to study, have at least the hope of accounting for a wide range of mental phenomena, and also helping us clarify our concepts of mental states and processes.

... continued

...Minds causation computation

- We conjecture that qualia arise out of the operation of certain forms of meta-managment, and when this is better understood many puzzles about consciousness will evaporate.
- The most interesting minds we know about are products of biological evolution and need to be understood as such.
- But that does not rule out creation of new sorts of minds with many similar capabilities, which are not the result of evolution. Some of them, for instance, could be disembodied minds that live forever in virtual worlds.
- As philosopher-engineers we'll have a good set of concepts for adopting what Dennett called the "design stance", directed at virtual machines and their innards. When we do that we shall not, as scientists, and philosophers, have any need for his "intentional stance", which is merely a fall-back for uninformed observers.

Future philosophical machines

If everything said above is correct, then

we can be sure some of our robots with meta-management will discover the strange differences between their internal VM processes and their physical bodies, and will re-discover much of philosophy of mind.

They may well wonder whether humans are really conscious.

Final session: 20 to 30 minutes

Possibly followed by ongoing discussions all week!

- **Free discussion of topics raised**
- **Identifying additional philosophical problems relevant to AI.**

Is there any feature of YOUR mind that cannot be accommodated in this framework?

Some references: A (fairly arbitrary) selection

BOOKS

J. Hospers, *Introduction to Philosophical Analysis* 4th ed., Prentice Hall, 1996.
(A GOOD GENERAL INTRODUCTION TO PHILOSOPHY)

David Lodge, *Thinks ....*
Secker and Warburg, 2001. (An entertaining novel which also provides a provocative introduction to some aspects of philosophy of mind and cognitive science).

P. Agre, *Computation and Human Experience*, Cambridge University Press, 1997.

J.S. Albus, *Brains, Behaviour and Robotics*, Byte Books, McGraw Hill, 1981,

M.A. Boden, *Artificial Intelligence and Natural Man*, Harvester Press, 1978, Second ed. MIT Press, 1986.

M.A. Boden, *The Creative Mind: Myths and Mechanisms*, Weidenfeld & Nicolson, 1990.

M.A. Boden (ed.), *The Philosophy of Artificial Intelligence*, Oxford Readings in Philosophy Series, Oxford University Press, 1990.

R.J. Brachman and H.J. Levesque (eds.), *Readings in Knowledge Representation*, Morgan Kaufmann.

K. Campbell, *Body and Mind*, Macmillan, 1970.

D.J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, 1996.

P.M. Churchland, *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind*, Cambridge, MA, MIT Press, 1984.

R. Chrisley (ed.), *Artificial Intelligence: Critical Concepts in Cognitive Science*, four volumes, Routledge, 2000.

R. Cummins, *Meaning and Mental Representation*, Cambridge, MA, MIT Press, 1997.

D.C. Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology*, MIT Press, Cambridge, MA, 1978.

D.C. Dennett, *Elbow Room: the varieties of free will worth wanting*, Oxford, The Clarendon Press, 1984.

D.C. Dennett, *Kinds of Minds*, Weidenfeld and Nicholson, 1996.

J.H. Fetzer (ed.), *Aspects of Artificial Intelligence*, Kluwer Academic, 1988.

J.H. Fetzer (ed.), *Epistemology and Cognition*, Kluwer Academic, 1990.

J.H. Fetzer *Artificial Intelligence: Its Scope and Limits*, Kluwer Academic, 1990.

J.A. Fodor *The language of thought* Harvester Press, 1976.

J.J Gibson, *The Ecological Approach to Visual Perception*, Lawrence Earlbaum Associates, 1986, (originally published 1979),

J. Glasgow, H. Narayanan, Chandrasekaran (eds.), *Diagrammatic Reasoning: Computational and Cognitive Perspectives*, AAAI Press, 1995.

J. Haugeland, *Artificial Intelligence: The Very Idea*, Bradford Books, MIT Press, 1985. (See the review by John McCarthy, below).

J. Haugeland (ed.), *Mind Design I*, Cambridge, MA, MIT Press, 1981.

J. Haugeland (ed.), *Mind Design II*, Cambridge, MA, MIT Press, 1996.

M. Hauser, *Wild Minds: What animals really think*, Penguin Books, 2001.

D.W. Hofstadter, *Godel, Escher, Bach: An Eternal Golden Braid*, Harvester Press, and Penguin books, 1979.

D.W. Hofstadter & D.C. Dennett (eds.), *The Mind's I: Fantasies and Reflections on Self and Soul*, Brighton, Harvester Press, 1981.

J. Kim, *Supervenience and Mind: Selected philosophical essays*, Cambridge University Press, Cambridge, 1993.

J. Kim, *Mind in a Physical World*, MIT Press, Cambridge, MA, 1998.

Bryan Magee, *Popper*, Fontana Modern Masters Series, 1973.

D. Marr, *Vision*, Freeman 1982,

M. L. Minsky, *The Society of Mind*, William Heinemann Ltd., London 1987,

A. Ortony, G.L. Clore & A. Collins, The Cognitive Structure of the Emotions, Cambridge University Press, 1988

J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000

H. Putnam, *Representation and Reality*, Cambridge, MA, MIT Press, 1988.

V. Pratt, *Thinking Machines - The Evolution of Artificial Intelligence*, Oxford, Basil Blackwell, 1987.

G. Ryle, *The Concept of Mind*, Hutchinson, 1949.

M. Scheutz, *The Missing Link: Implementation and Realization of Computations in Computer and Cognitive Science*, PhD Thesis, Indiana University, (University of Michigan Microfiche), 1999.

J. Searle, *The Rediscovery of Mind*, Cambridge, Massachusetts, MIT Press, 1992.

A. Sloman, *The computer revolution in philosophy: philosophy science and models of mind*, Harvester press, 1978. (Out of print but available from here: http://www.cs.bham.ac.uk/˜axs/cpr.html it is being scanned in.)

B.C. Smith, *On the Origin of Objects*, MIT Press, 1996.

B.C. Smith, *The Age of Significance*, seven volumes, MIT Press (forthcoming).

N.A. Stillings, et al. (eds.), *Cognitive Science: An Introduction*, 2nd edition, Cambridge, MA, MIT Press, 1995.

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

ARTICLES

M. Alliksaar, ‘Metabolism in A-Life: Reply to Boden’, in *The British Journal for the Philosophy of Science*, vol 52, pp. 131–135, 2001

J.A. Barnden & M. Lee, ‘An implemented context system that combines belief reasoning, metaphor-based reasoning and uncertainty handling, in *Proceedings of the 2nd International and Interdisciplinary Conference on Modelling and Using Context*. Eds. P, Bouquet, P. Brezillon, L. Serafini, LNAI 1688, Springer, pp. 28 – 41., 1999

M.A. Boden, ‘Is metabolism necessary?’, in *The British Journal for the Philosophy of Science*, vol 50, pp. 231–248, 2001

A. Botterell ‘Conceiving what is not there’, *Journal of Consciousness Studies*, vol 8, no 8, pp 21–42, 2001

R. A. Brooks, ‘Intelligence without representation’, in *Artificial Intelligence*, 47, pp. 139–159, 1991, (Also in his online paper directory – see below)

N. Chomsky, A Review of B. F. Skinner’s Verbal Behavior. *Language* 35(1):26-58, 1959 (available at http://cogprints.soton.ac.uk/documents/disk0/00/00/11/48/)

R. Cooper and T. Shallice, ‘Contention scheduling and the control of routine activities’, in *Cognitive Neuropsychology*, 17, 4, pp. 297–338, 2000,

D. Davidson, ‘Mental Events’, first published 1970, reprinted in: *Essays on Action and Events* (Oxford: Oxford University Press, 1980).

D. Kirsh, ‘Today the earwig, tomorrow man?’, *Artificial Intelligence* vol 47, nos 1-3, pp. 161–184, 1991

J. McCarthy, ‘Ascribing mental qualities to machines’, in *Philosophical Perspectives in Artificial Intelligence*, Ed. M. Ringle, Humanities Press, pp.161–195 1979, (Also at http://www-formal.stanford.edu/jmc/ascribing/ascribing.html),

J. McCarthy, Online book reviews, http://www-formal.stanford.edu/jmc/reviews/

J. McCarthy, ‘Review of ”Artificial Intelligence: The Very Idea”, (Haugeland 1985), available at http://www-formal.stanford.edu/jmc/reviews/haugeland.html

J. McCarthy and P.J. Hayes, Some philosophical problems from the standpoint of AI, in *Machine Intelligence 4*, Eds. B. Meltzer and D. Michie, Edinburgh University Press, 1969, (Accessible as http://www-formal.stanford.edu/jmc/mcchay69/mcchay69.html)

M. L. Minsky, ‘Steps towards Artificial Intelligence’, in *Computers and Thought*, Eds. E.A. Feigenbaum and J. Feldman, McGraw-Hill, New York, pp. 406–450, 1963, (Book reprinted 1995, MIT Press: Includes several very influential articles),

M.L. Minsky, ‘Future of AI Technology’, http://www.media.mit.edu/people/minsky/papers/CausalDiversity.html, 1997 Original version in Toshiba Review, Vol.47, No.7, July 1992.

John R Searle, ‘Minds Brains and Programs’, *The Behavioral and Brain Sciences*, 3, 3, 1980, (With commentaries and reply by Searle)

H. A. Simon, ‘Motivational and emotional controls of cognition’, 1967, Reprinted in *Models of Thought*, Yale University Press, 29–38, 1979

A. Sloman, ‘The emperor’s real mind’, Review of Roger Penrose’s *The Emperor’s new Mind: Concerning Computers Minds and the Laws of Physics*, in *Artificial Intelligence*, 56, pp. 355–396, 1992

A. Sloman, ‘The mind as a control system’, in *Philosophy and the Cognitive Sciences*, Ed. C. Hookway and D. Peterson, Cambridge University Press, pp. 69–110, 1993,

A. Sloman, ‘Towards a general theory of representations’, in *Forms of representation: an interdisciplinary theme for cognitive science*, Ed. D.M.Peterson, Intellect Books, Exeter, U.K., pp. 118–140, 1996a,

A. Sloman, ‘Actual Possibilities’, in *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR ‘96)*, Eds. L.C. Aiello and S.C. Shapiro, Morgan Kaufmann Publishers, pp. 627–638, 1996b,

A. Sloman, ‘Architectural requirements for human-like agents both natural and artificial. (What sorts of machines can love?)’, in *Human Cognition And Social Agent Technology*, Ed. K. Dautenhahn, Advances in Consciousness Research, John Benjamins, pp. 163–195, 2000

A. Sloman, ‘Interacting trajectories in design space and niche space: A philosopher speculates about evolution’, in *Parallel Problem Solving from Nature – PPSN VI*, Eds. M.Schoenauer *et al.*, Springer, LNCS No 1917, pp. 3–16, 2000.

R. Sun, E. Merrill & Todd Peterson ‘From implicit skills to explicit knowledge: a bottom-up model of skill learning’, in *Cognitive Science*, 25,2, pp. 203–244 2001

A.M. Turing, ‘Computing machinery and intelligence’, *Mind*, 59, pp. 433–460, 1950, (reprinted in E.A. Feigenbaum and J. Feldman (eds) Computers and Thought McGraw-Hill, New York, 1963, 11–35).

T.J. van Gelder, ‘The dynamical hypothesis in cognitive science’ in *Behavioral and Brain Sciences*, 21, 1–14. 1998 (Followed by critical comments and replies by author.)

SOME WEB SITES

AAAI web site on AI, maintained by Jon Glick http://www.aaai.org/aitopics

AAAI web site on AI, philosophy section http://www.aaai.org/aitopics/html/phil.html

Rodney Brooks, home page and online papers, http://www.ai.mit.edu/people/brooks/brooks.html

David Chalmers, Consciousness bibliography, http://www.u.arizona.edu/˜chalmers/

Fred Dretske First person warrant: comments on Siewert’s *The significance of consciousness* http://psyche.cs.monash.edu.au/v7/psyche-7-11-dretske.html

Encyclopaedia Britannica web site on metaphysics:
http://www.britannica.com/bcom/eb/article/4/0,5716,115524+6,00.html

Immanuel Kant, *Critique of Pure Reason* online http://www.arts.cuhk.edu.hk/Philosophy/Kant/cpr/

Peter King's philosophy pages: http://users.ox.ac.uk/~worc0337

John McCarthy's web site: http://www-formal.stanford.edu/jmc

Marvin Minsky's web site: http://www.media.mit.edu/ minsky/

Nils Nilsson's Web site: http://www.robotics.stanford.edu/users/nilsson/ including his Teleoreactive systems page: http://www.robotics.stanford.edu/users/nilsson/trweb/tr.html

Judea Pearl's web site: http://bayes.cs.ucla.edu/jp_home.html

A. Sloman, 'Architectures for intelligent language users (How to turn philosophers of mind into engineers and vice versa)', Lecture slides for course at *ESSLLI-2000, Twelfth European Summer School in Logic, Language and Information*, August 2000. http://www.cs.bham.ac.uk/~axs/esslli/esslli.slides.ps (or .pdf)

A. Sloman and others *The Cognition and Affect Project*, University of Birmingham
http://www.cs.bham.ac.uk/research/cogaff/

THERE ARE MANY, MANY MORE: CHECK WWW.GOOGLE.COM AND OTHER SEARCH ENGINES.

ACKNOWLEDGEMENTS

This work is partly funded by grant F/94/BW from the Leverhulme Trust, for research on 'Evolvable virtual information processing architectures for human-like minds'.

The ideas presented here were inspired by the work of many philosophers and AI researchers, including Margaret Boden, Dan Dennett, Brian Cantwell Smith, Pat Hayes, John McCarthy, Marvin Minsky, Herbert Simon.

Our own ideas were developed in collaboration with past and present colleagues in the Cognition and Affect Project in the School of Computer Science, The University of Birmingham http://www.cs.bham.ac.uk/

Especially Steve Allen, John Barnden, Luc Beaudoin, Catriona Kennedy, Brian Logan, Donald Peterson, Riccardo Poli, Ian Wright.

We have also learnt much from the work of colleagues in the School of Psychology, including Glyn Humphreys, Jane Riddoch and Alan Wing.

David Rose (Surrey University) made several useful comments.

Questions and comments raised at the tutorial led to some changes in the revised version of the slides.