ARTIFICIAL INTELLIGENCE AND PHILOSOPHY

How AI (including robotics) relates to philosophy and in some ways

Improves on Philosophy

Aaron Sloman

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

School of Computer Science

The University of Birmingham

Presented to MSc students and Undergraduates University of Birmingham
Accessible later here

http://www.cs.bham.ac.uk/research/cogaff/talks/#talk109

(Several other talks are also relevant)

Also relevant

- The Computer Revolution in Philosophy (1978)
  http://www.cs.bham.ac.uk/research/cogaff/crp/
- What is science? Can there be a science of mind?
  http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk18
- The Meta-Morphogenesis project
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html

CONTENTS

We shall touch on the following topics – all huge

- What is philosophy?
- What is science?
- What is engineering?
- How does philosophy relate to science, engineering and mathematics?
- What is AI?
- How does philosophy relate to AI?
- Some examples
- NOTE: there are lots more talks and discussions on my web pages, e.g.
  `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/AREADME.html`

NOTE: my computer uses Linux

WHAT IS PHILOSOPHY?

Philosophy asks the questions nobody knows how to answer!

The most general of all forms of enquiry, with many more specific spin-offs.

PHILOSOPHY INVESTIGATES:

- The most general questions about what exists:
  - Metaphysics and ontology:

    Attempt to categorise the most general forms of reality and possibly to explain why reality is like that.
    E.g. Can mind exist independently of matter?
- The most general questions about questions and possible answers:
  - Epistemology:

    an attempt to characterise the nature of knowledge and to identify the kinds of knowledge that are possible and the ways of acquiring knowledge.
  - Theory of meaning:

    An attempt to clarify the nature of meaning and how it differs from nonsense.
- The most general questions about what ought to exist, or ought not to exist:
  - Ethics (moral philosophy) and aesthetics

    an attempt to distinguish what is good and what is bad, including what is good or bad in art.
    Meta-ethics investigates the nature of ethics.
- **ALSO** Philosophy of X:

  Where X is science, mathematics, history, art, biology, physics, medicine ....

Contrast a naive popular view of philosophy: a study of “the meaning of life”?

WHAT IS SCIENCE?

1. A common answer:

   Science is A search for the laws of nature???

   We can improve on that answer!

   SEE:

   `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk18`

2. Science is an attempt to find out

   – what sorts of things can exist,

   – what sorts of processes can exist,

   – what mechanisms an explain how they come to exist

CRAFT (practical skills, know-how) leads to SCIENCE

SCIENCE is used by ENGINEERING (knowledge based craft)

But there are not sharp boundaries.

More specific areas of philosophy

Besides the very general branches of philosophy there are many sub-branches which combine the above three philosophical studies in focusing on a particular form of human activity: Philosophy of X.

Examples include:

- Philosophy of mind
- Philosophy of mathematics
- Philosophy of language
- Philosophy of science
- Philosophy of history
- Philosophy of economics
- Philosophy of biology
- Philosophy of psychology
- Philosophy of literature
- Philosophy of education
- Philosophy of politics
- Philosophy of computation
- Philosophy of music
- Philosophy of sport (e.g. what makes a competition fair?)
- Philosophy of ....

Philosophy of mind is close to AI

Philosophy of mind has several different aspects, all relevant to AI:

- Metaphysical and ontological topics (what exists)

  Questions about the nature of mind and the relation between mind and body, e.g. whether and how mental events can cause physical events or *vice versa*.

  Compare virtual machines in computers

- Epistemological topics (theory of knowledge)

  Questions about whether we can know about the minds of others, and how we can acquire such knowledge.

  More subtle questions about what we can and cannot know about our own minds: e.g. do you know which rules of grammar you use?

  Compare: what can different sorts of robot know?

- Conceptual analysis (what do we mean by X?)

  Analysis of the concepts we use in talking about our mental states and processes, e.g. ‘perceive’, ‘desire’, ‘think’, ‘plan’, ‘decide’, ‘enjoy’, ‘conscious’, ‘experience’ ...

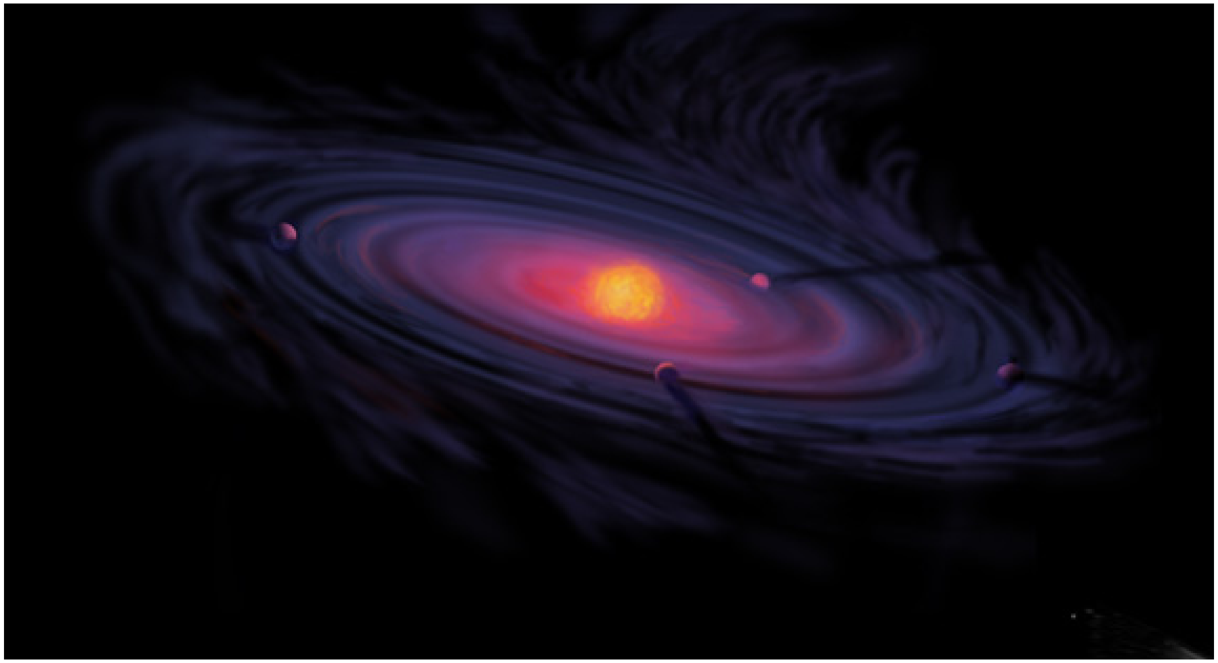
  Important for clarifying terms used to describe a robot

- Methodology (e.g. philosophy of psychology)

  Investigation, comparison and evaluation of the various ways of studying human (and animal) minds, including the methods of psychology, neuroscience, social science, linguistics, philosophy, and AI.

A new synthesis Philosophy AI and Biology

The meta-morphogenesis project.



How can a cloud of dust give birth to a planet full of living things as diverse as life on Earth?

What sorts of answers are possible?

What we've learnt about computing and AI in the last half century gives us new ways of thinking about this that were not available to Darwin, and are still ignored by most philosophers.

AI extends philosophy of mind

AI can be seen as an extension of philosophy of mind:

- Philosophers ask what necessary and sufficient conditions for minds to exist are: as if there could be only one kind of mind – AI investigates varied designs for different minds.
- We can survey different possible kinds of minds by asking how we could design and implement them. (So far AI has produced only very simple examples.)
- We can clarify the relationship between mind and body by treating it as a special case of another relationship that we understand better: the relationship between virtual machines (running programs) and physical machines (computers).

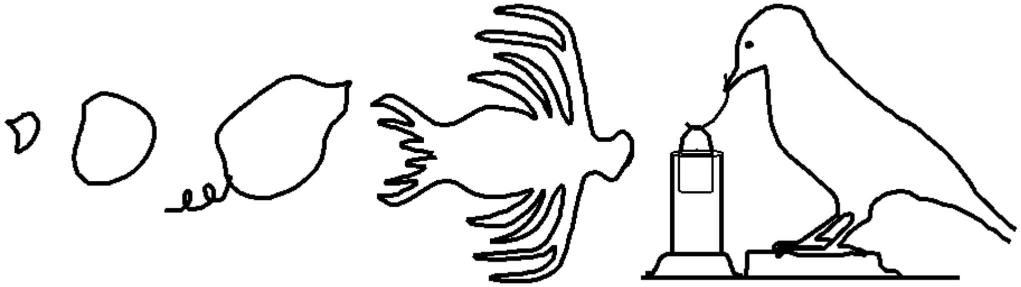
  (Virtual machines have many of the features of minds that have puzzled philosophers.)
- We can explore different *architectures* for minds, and see which sorts of *concepts* are appropriate for describing the different sorts of minds

  (e.g. concepts like ‘perception’, ‘thinking’, ‘emotion’, ‘belief’, ‘pleasure’, ‘consciousness’.)
- We can address the ‘problem of other minds’ (how do we know anything about another mind?) by exploring architectures for agents that need to be able to think about and communicate with other agents.

  (Different kinds of awareness of other agents in predators, prey, social animals, etc.)
- By attempting to design *working* models of human minds, and noticing how our programs are inadequate, we discover some unobvious facets of our own minds, and some unobvious requirements (e.g. for perception, learning, reasoning).

A short history of life

All organisms are information-processors but the information to be processed, the uses of the information, the processing mechanisms, and the architectures used, have all varied enormously, between the earliest microbes and sophisticated modern animals.



Evolution (a) isn't a uniform process in which natural selection uses random chemical changes, and (b) it isn't continuous (e.g. molecular changes are inherently discrete):

Many different things can influence evolution, including

- What's in the environment

  Physical structures and processes, commonalities and differences across spaces and times.

  Other information-processing systems – prey, predators, competitors, conspecifics, offspring, ...

- Products of previous evolution (evolved internal and external niche features)

  Previous physical developments and previous computational developments in the species, or in other species (prey, predators, competitors, symbionts, ...)

  These provide new requirements (e.g. new requirements to control newly articulated body parts, or to get the most benefit from new sensory mechanisms, or to hide from flying predators.)

  They also provide new opportunities: new platforms for further development, and new constraints.

Changing environmental influences on evolution

Types of environment with different information-processing requirements

What information processing mechanisms could be useful in the following epochs?

- Microbes in a chemical soup
- Microbes in a soup with detectable chemical gradients
- Soup plus some stable structures (places with good stuff, bad stuff, obstacles, supports, shelters)
- Things in the environment that have to be manipulated to be eaten (e.g. disassembled)
- Organisms with controllable "effectors" - for moving, eating, building, fighting...
  (products of previous evolution provide new opportunities for information-based control)
- Food/prey/predators/mates with detectable aromas (more chemical information bearers)
- Environments with food that tries to escape (prey).
- Organisms in environments with things that try to eat them (predators).
- Environments with and without places to hide from predators
- Environments with and without places for prey to hide
- Prey, predators, collaborators, whose behaviour can reveal intentions, interests, knowledge...
- Mates with preferences
- Competitors for food and mates
- Collaborators that need, or can supply, information.
- and so on .....

If we analyse changes in information processing requirements (niches) and designs, we can construct a collection of "dependency trees" showing possible evolutionary and developmental trajectories? E.g. see:
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/evolution-info-transitions.html`
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-requirements.html`

We need deep theories of types of requirement and how they interact and change

Philosophy needs AI and AI needs philosophy

- AI needs philosophy to help clarify
  - its goals: e.g. what is the study of intelligence? what are intelligent machines?,
  - the concepts it uses,
  - the kinds of knowledge and ontology a machine needs to interact with us,
  - some methodological issues: e.g. how can AI theories be tested? Are the goals of AI achievable?

- Philosophy needs AI
  - To provide a new context for old philosophical questions.
    E.g. 'What can be known?' becomes a more focused question in the context of different specific sorts of machines that can perceive, infer, learn, ...
  - To provide a host of new phenomena to investigate, partly to clarify old philosophical concepts and theories. E.g.
    - New kinds of machines: information processing machines
    - New kinds of representation, inference, communication
    - New examples of physical non-physical interaction: virtual machines and computers.
    - New sorts of virtual machine architectures

New examples help to refute bad old theories and to clarify old concepts — good and bad.

(See the 'Philosophical encounter' in IJCAI 1995 (Minsky, McCarthy and Sloman))

WHAT IS ENGINEERING?

According to Wikipedia:

Engineering

(from Latin ingenium, meaning “cleverness” and ingeniare, meaning “to contrive, devise”)

is

> the application of scientific, economic, social, and practical knowledge in order to invent, design, build, maintain, research, and improve structures, machines, devices, systems, materials and processes.

1. Doing science is attempting to understand the world
2. Doing engineering is attempting to change the world

   What about improving/changing minds?
   Is education a form of engineering?

   In order to improve something it is useful to know how it works.

   What about building new sorts of minds?

What’s left for philosophy?

Often neither current science nor current engineering is sufficiently well developed to provide the required answers to questions and solutions to problems.

Philosophy is often at its best as feeding into science (and engineering) by

– posing new questions that are not yet answerable

– helping to clarify questions

– helping to evaluate what other disciplines do

And challenging confused or question-begging, circular answers.

(e.g. challenging attempts to use religious beliefs to answer questions instead of studying the world to find out the facts.)

Philosophy can help to clarify goals: why should we do that?.

It challenges assumptions about what we ought to do, how we ought to live, what we ought not to do.

Both science and philosophy challenge answers that come from “authorities”, but they do so in different ways.

Science and philosophy

Science:

finding out facts about the world, and testing claims about the facts

Philosophy: getting clear about

– what we mean,

– what questions we are asking,

– what our goals are, and whether they are achievable,

– what forms of evidence are relevant to what questions, and

– what sorts of justifications are appropriate for choosing goals, values, means to ends.

A risk in philosophy

Danger:

professional philosophy can lead mutual navel-gazing.

Sometimes philosophers read only acknowledged experts in philosophy and then their discussions risk turning into “disconnected chatter”.

For example: trying to understand what minds are, by discussing only what other philosophers have said minds are,

Ignoring

– detailed attempts to study minds (psychology, biology, neuroscience)

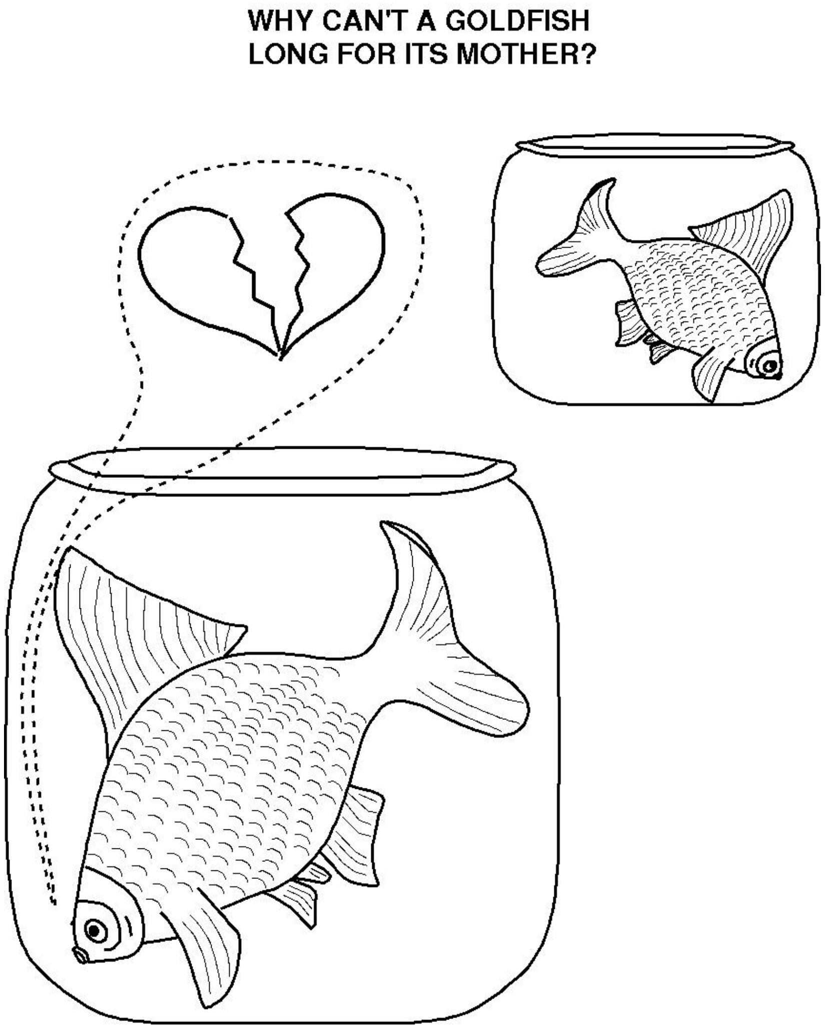
– detailed attempts to build minds (artificial intelligence/robotics)

(Ignoring engineering that’s relevant to philosophy)

Often attempting to replicate things is one of the best ways to try to understand them:

philosophical engineering → **AI**

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...??
- Because it lacks an “emotion sub-system”

No, because:

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

For example, there is no reason to believe that a goldfish has any concept of a mother in general or of its mother in particular or that it can conceive of the possibility of being in the vicinity of its mother, or that thinking about the fact that that possibility is not realised could produce changes amounting to “longing for”, or being sad, etc.

Compare current “emotional” robots.

A Core Question in Philosophy of Mind

What kind of thing is a mind?

- Minds (or what some people call ‘souls’) seem to be intangible and to have totally different properties from material objects.
- For instance physical objects have weight, size (e.g. diameter) and shape, yet thoughts, feelings, intentions, have none of those properties.
- We can discover the physical properties of people by observing them, measuring them in various ways, and if necessary cutting them open, but we cannot discover their mental states and feelings like that: we mostly depend on them to tell us, and we have a special way of becoming aware of our own.
- Such facts have led some people to question whether mental phenomena can really exist in our universe.
- However, we can get a better understanding of these matters if we realise that not only minds have this relationship to matter: reality is composed of entities at multiple levels of abstraction with different properties.

How to think about non-physical levels in reality

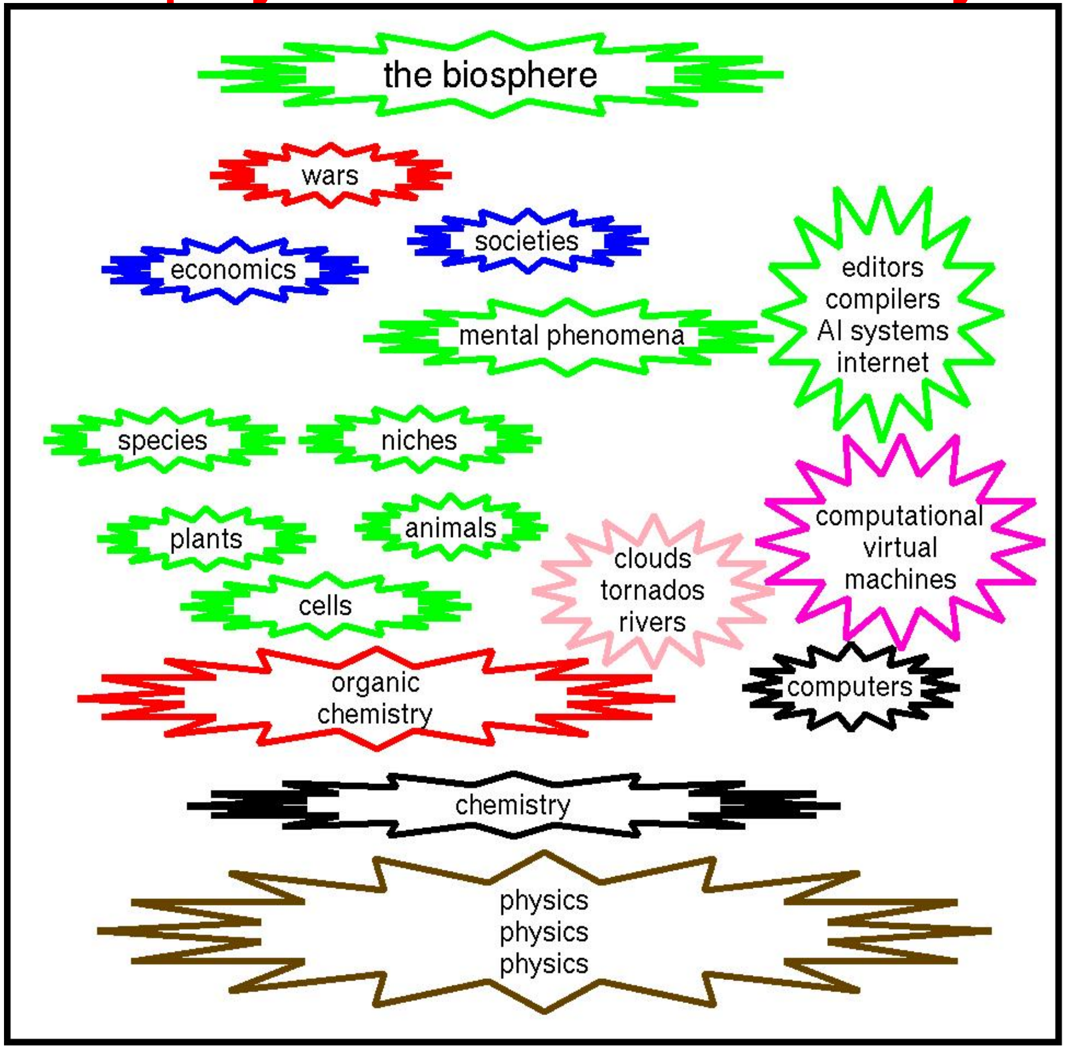
Some philosophers think only physical things can be real.

But there are many non-physical objects, properties, relations, structures, mechanisms, states, events, processes and causal interactions.

E.g. poverty can cause crime.

They are all ultimately implemented in physical systems, as computational virtual machines are, e.g. the Java VM, the linux VM.

Physical sciences also study layers in reality. E.g. chemistry is implemented in physics. Nobody knows how many levels of virtual machines physicists will eventually discover.



See this introduction to Virtual-Machine functionalism:
`http://tinyurl.com/CogMisc/vm-functionalism.html`

DIFFERENT VIEWS OF MIND

OLDER APPROACHES:

- A ghost in a machine (dualism)
  - With causal connections both ways: Interactionism
  - With causal connections only one way: Epiphenomenalism
  - With no causal connections: Pre-established harmony
- Mind-brain identity (e.g. the double-aspect theory)
- Behaviourism (mind defined by input-output relations)
- Social/political models of mind
- Mechanical models (e.g. levers, steam engines)
- Electrical models (old telephone exchanges)

PROBLEMS WITH OLDER APPROACHES

- Some lack explanatory power (ghost in the machine)
- Some are circular (Social/Political models of mind)
- Some offer explanations that are too crude to explain fine detail and do not generalise (e.g. mechanical and electrical models)

AI provides tools and concepts for developing new rich and precise theories which don’t merely describe some overall structure of mind or mind-body relation, but can show how minds work.

Is there a ghost in the machine?

In 1949, the philosopher Gilbert Ryle wrote a very influential book called ‘The Concept of Mind’ criticising the theory of the ghost in the machine. (It is well worth reading.)

But in those days they did not know much about how to make ghosts in machines.
Now we know how to put a functioning virtual machine (e.g. an operating system or spelling checker) inside a physical machine,

If there is a ghost in the machine it requires sophisticated information-processing capabilities to do what minds do.

I.e. there must be a machine in the ghost – an information processing virtual machine.

Only a virtual machine can have sufficient flexibility and power
(as evolution discovered before we did.)
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#cogsci09

(We need to investigate different sorts of virtual machine.)

What AI adds

AI enables philosophy to take account of information-processing virtual machines

The Birmingham Cognition and Affect project attempts to develop a new philosophy of mind:

Virtual machine functionalism

See

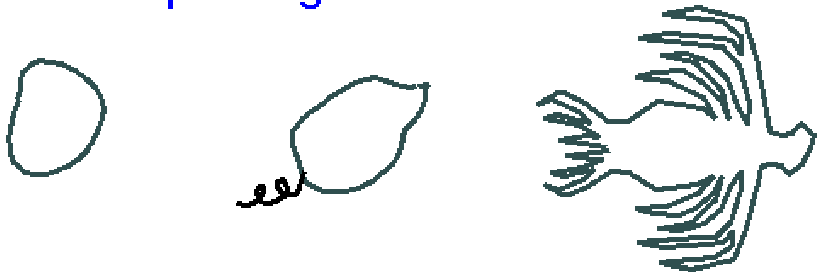
```
http://www.cs.bham.ac.uk/research/cogaff/talks/#super
http://www.cs.bham.ac.uk/research/cogaff/talks/#inf
```

Mental concepts are defined in terms of states and processes in Virtual machines with complex information processing architectures.

Organisms process information

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

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



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

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

NOTE:

We use "information" in the everyday sense, which involves notions like "referring to something", "being about something", "having meaning", not the Shannon/Weaver technical sense, which is a purely syntactic notion.
Compare Jane Austen's use of the word in *Pride and Prejudice*
`http://tinyurl.com/CogMisc/austen-info.html`

Resist the urge to ask for a definition of "information"

Compare "energy" – the concept has grown much since the time of Newton. Did he understand what energy is?

Instead of defining "information" we need to analyse the following:

- the variety of types of information there are,
- the kinds of forms they can take,
- the means of acquiring information,
- the means of manipulating information,
- the means of storing information,
- the means of communicating information,
- the purposes for which information can be used,
- the variety of ways of using information.

As we learn more about such things, our concept of "information" grows deeper and richer.

Like many deep concepts in science, it is *implicitly* defined by its role in our theories and our designs for working systems.

For more on this see:

http://www.cs.bham.ac.uk/research/projects/cogaff/misc/whats-information.html

Things you can do with information

A partial analysis to illustrate the above:

- You can react immediately (information can trigger immediate action, either external or internal)
- You can do segmenting, clustering labelling of components within a complex information structure (i.e. do parsing.)
- You can interpret one entity as referring to something else.
- You can try to derive new information from old (e.g. what caused this? what else is there? what might happen next? can I benefit from this?)
- You can store store information for future use (and possibly modify it later)
- You can consider alternative possibilities, e.g. in planning.
- If you can interpret it as as containing instructions, you can obey them, e.g. carrying out a plan.
- You can observe the process of doing all the above and derive new information from it (self-monitoring, meta-management).
- You can communicate it to others (or to yourself later)
- You can check it for consistency, either internal or external

... All of this can be done using different forms of representation for different purposes.

What an organism or machine can do with information depends on its architecture

Not just its physical architecture – its information processing architecture.

This may be a virtual machine, like

- a chess virtual machine
- a word processor
- a spreadsheet
- an operating system (linux, solaris, windows)
- a compiler
- most of the internet

What is an architecture?

AI used to be mainly about algorithms and representations.

Increasingly, during the 1990s and onward it has been concerned with the study of architectures.

An architecture includes:

– forms of representation,
– algorithms,
– concurrently processing sub-systems,
– connections between them.

Note: Some of the sub-systems may themselves have complex architectures.

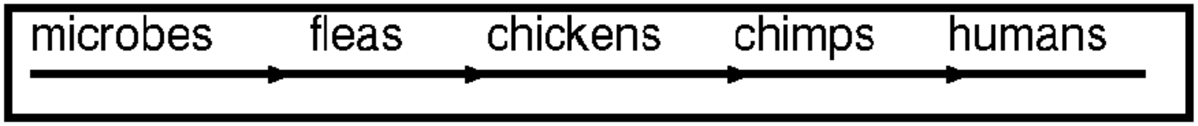
We need to understand the space of information processing architectures and the states and processes they can support, including the varieties of types of mental states and processes.

Which architectures can support human-like emotions?

There’s No Unique Correct Architecture

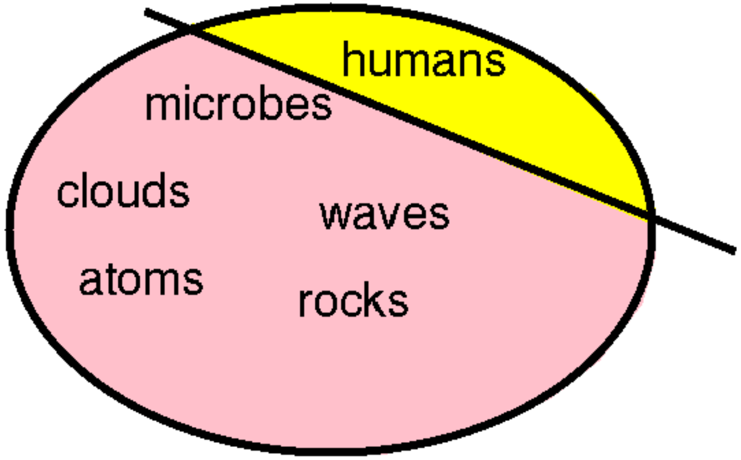
Some tempting wrong ways to think about consciousness:

1. There's no continuum from non-conscious to fully conscious beings



microbes → fleas → chickens → chimps → humans

2. It's not a dichotomy either



Both ‘smooth variation’ and a single discontinuity are poor models for kinds of variation in biology, chemistry, designs for information processing systems, ...

We need a better view of the space of possibilities

There are many different types of designs, and many ways in which designs can vary.

Some variations are continuous

(getting bigger, faster, heavier, etc.).

Some variations are discontinuous:

- duplicating a structure,
- adding a new connection between existing structures,
- replacing a component with another,
- extending a plan.
- adding a new control mechanism

Many biological changes are discontinuous

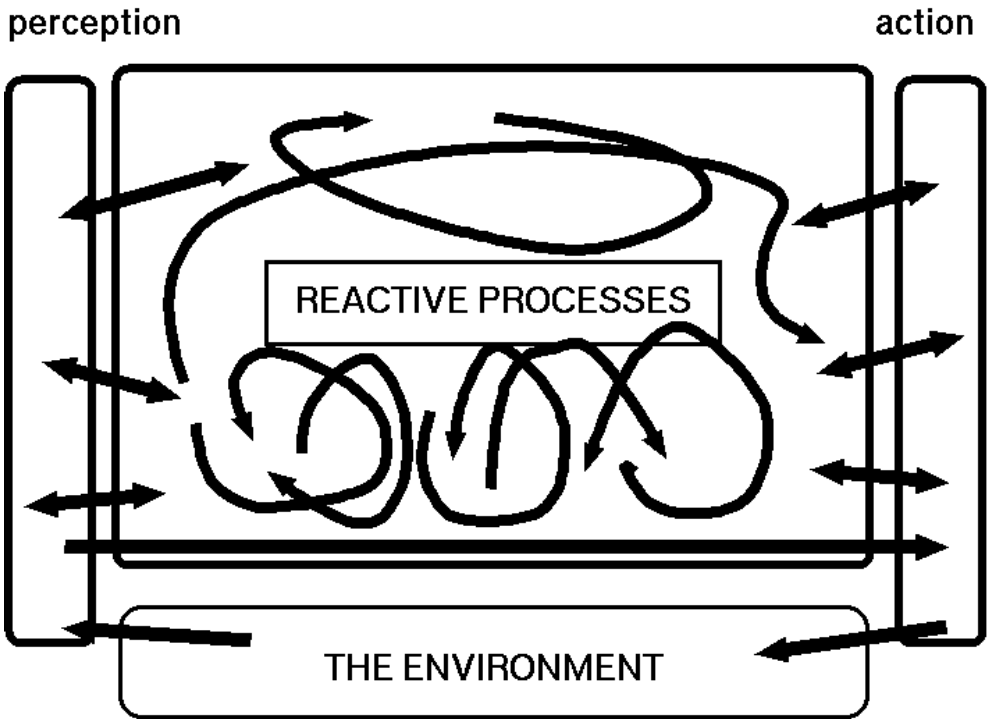
Discontinuities can be big or small.

In particular, changes of kind as well as degree occur in all of:

- evolution,
- development of an embryo from an egg,
- development of a child's mind

A simple (insect-like) architecture

A reactive system does not construct descriptions of possible futures evaluate them and then choose one. It simply reacts (internally or externally).



An adaptive system with reactive mechanisms can be a very successful biological machine. Some purely reactive species also have a social architecture, e.g. ants, termites, and other insects.

Features of reactive organisms

The main feature of reactive systems is that they lack the ability to represent and reason about non-existent phenomena (e.g. future possible actions), the core ability of deliberative systems, explained below.

Reactive systems need not be “stateless”: some internal reactions can change internal states, and that can influence future reactions.

In particular, reactive systems may be adaptive: e.g. trainable neural nets, which adapt as a result of positive or negative reinforcement.

Some reactions will produce external behaviour. Others will merely produce internal changes.

Internal reactions may form loops.

An interesting special case are teleo-reactive systems, described by Nils Nilsson (`http://robotics.stanford.edu/`)

In principle a reactive system can produce any external behaviour that more sophisticated systems can produce: but possibly requiring a larger memory for pre-stored reactive behaviours than could fit into the whole universe. Evolution seems to have discovered the advantages of deliberative capabilities.

Some people do not believe biological evolution occurred.

It’s strange to think that some people think their God could not produce biological evolution even though human software engineers can produce evolutionary processes in computers.

See this discussion on “Intelligent Design” `http://www.cs.bham.ac.uk/~axs/id`

“Consciousness” in reactive organisms

Is a fly conscious of the hand swooping down to kill it?

Insects perceive things in their environment, and behave accordingly.

However, it is not clear whether their perceptual mechanisms produce information states between perception and action usable in different ways in combination with different sorts of information.

(Compare the different ways you can use information that a table is in the room.)

Rather, it seems that their sensory inputs directly drive action-control signals, though possibly after transformations which may reduce dimensionality, as in simple feed-forward neural nets.

There may be exceptions: e.g. bees get information which can be used either to control their own behaviour or to generate “messages” that influence the behaviour of others.

Typically a purely reactive system does not use information with the same type of flexibility as a deliberative system which can consider non-existent possibilities.

They also lack self-awareness, self-categorising abilities. A fly that sees an approaching hand probably does not know that it sees — it lacks meta-management mechanisms, described later.

Demonstrations of reactive architectures

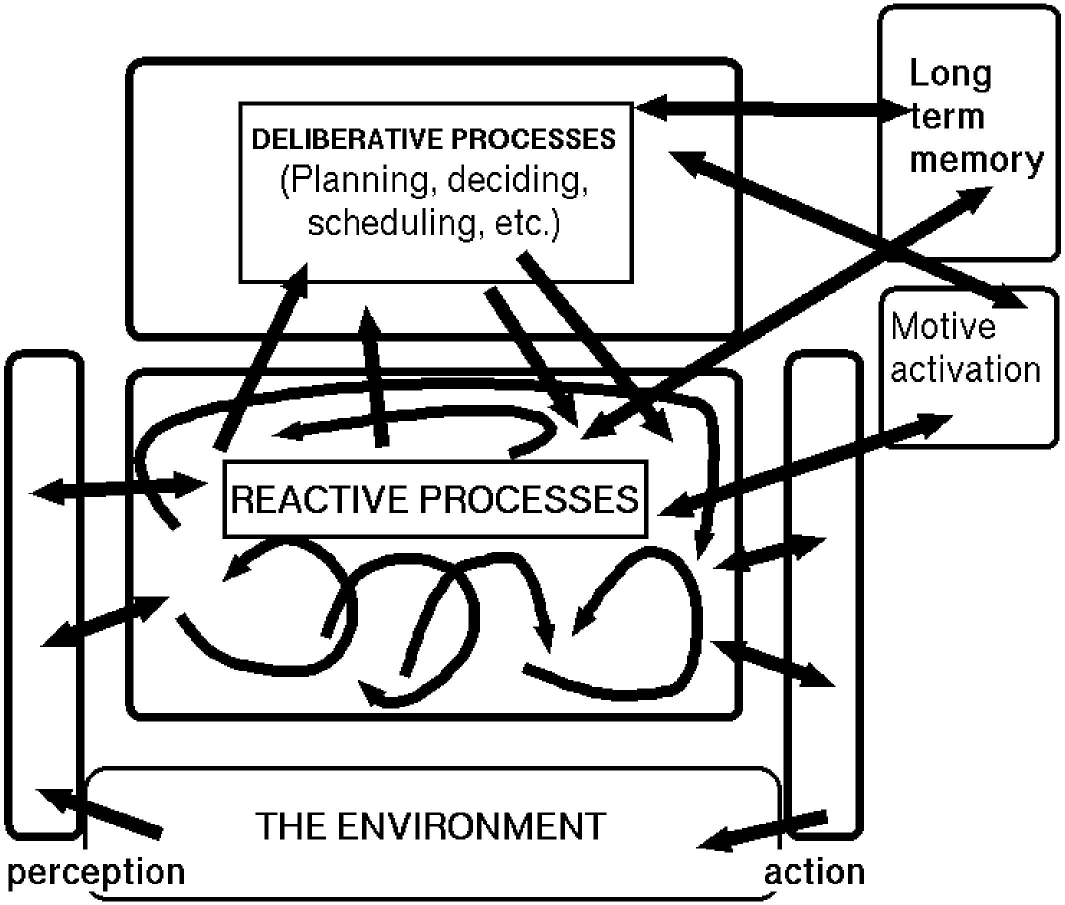
Sheepdog
‘Emotional’ agents
(Others if there is time.)

The demos are available in non-interactive mode (as movies) here

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

Note: the demos also illustrate causation in virtual machines.

Sometimes the ability to plan is useful



Deliberative mechanisms provide the ability to represent possibilities (e.g. possible actions, possible explanations for what is perceived).

Much, but not all, early symbolic AI was concerned with deliberative systems (planners, problem-solvers, parsers, theorem-provers).

Deliberative Demos

- SHRDLU (pop11 gblocks)
- The ‘hybrid’ sheepdog that interleaves planning, plan execution, and reactive behaviours.

The demos are available in non-interactive mode (as movies) here

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

Note: the demos illustrate causation in virtual machines.

Deliberative mechanisms

These differ in various ways:

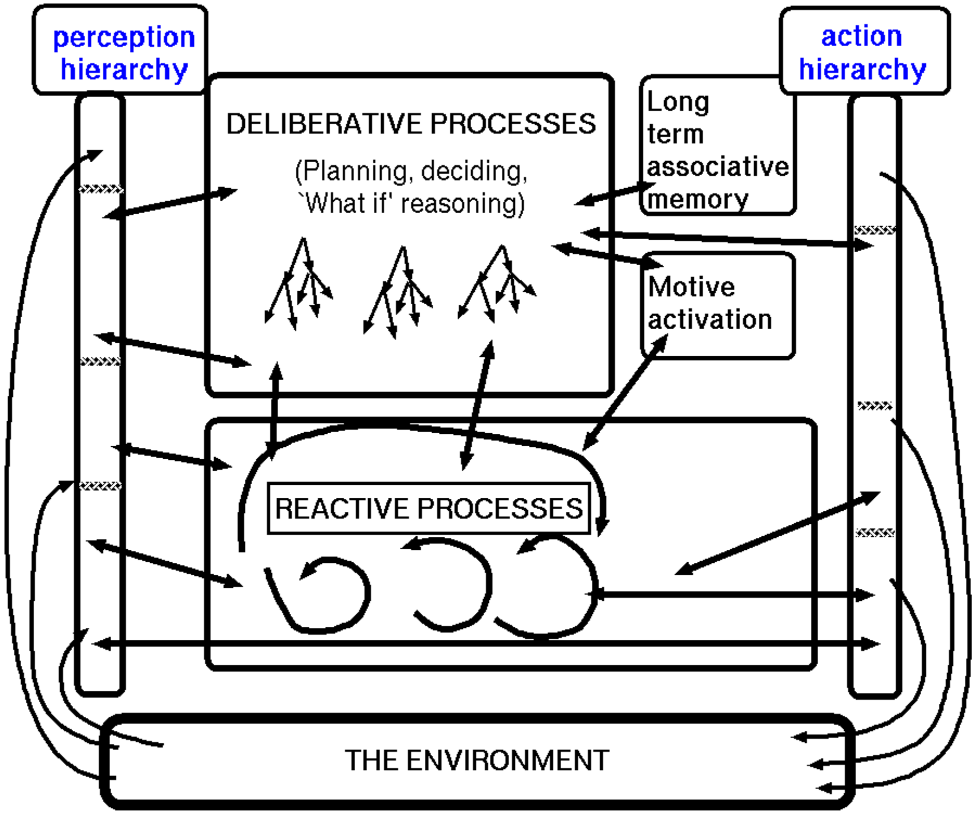
- the forms of representations (often data-structures in virtual machines)
- the variety of forms available (e.g. logical, pictorial, activation vectors)
- the algorithms/mechanisms available for manipulating representations
- the number of possibilities that can be represented simultaneously
- the depth of ‘look-ahead’ in planning
- the ability to represent future, past, or remote present objects or events
- the ability to represent possible actions of other agents
- the ability to represent mental states of others (linked to meta-management, below).
- the ability to represent abstract entities (numbers, rules, proofs)
- the ability to learn, in various ways

Some deliberative capabilities require the ability to learn new abstract associations, e.g. between situations and possible actions, between actions and possible effects

For more details see

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604`
Requirements for a Fully Deliberative Architecture

Evolutionary pressures on perceptual and action mechanisms for deliberative agents

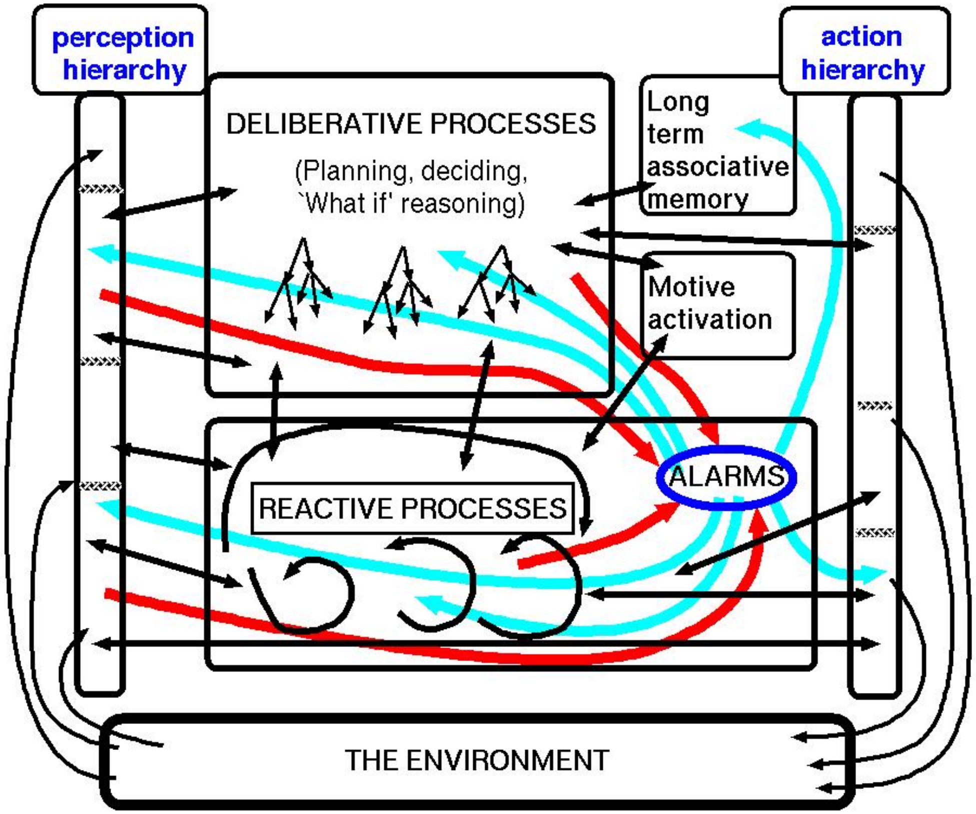


New levels of perceptual abstraction (e.g. perceiving object types, abstract affordances), and support for high-level motor commands (e.g. “walk to tree”, “grasp berry”) might evolve to meet deliberative needs – hence taller perception and action towers in the diagram.

‘Multi-window’ perception and action, vs ‘peephole’ perception and action (in many architectures).

A deliberative system may need an alarm mechanism

This is one possible way to represent the diversity of “alarm” functions (rapid-reaction functions?) in a complex information processing architecture.



For some purposes it may be useful to separate the “alarm” subsystem into different systems, operating in parallel and performing different monitoring tasks (and a mechanism for dealing rapidly with conflicts between them?)

Inputs to an alarm mechanism may come from anywhere in the system, and outputs may go to anywhere in the system.

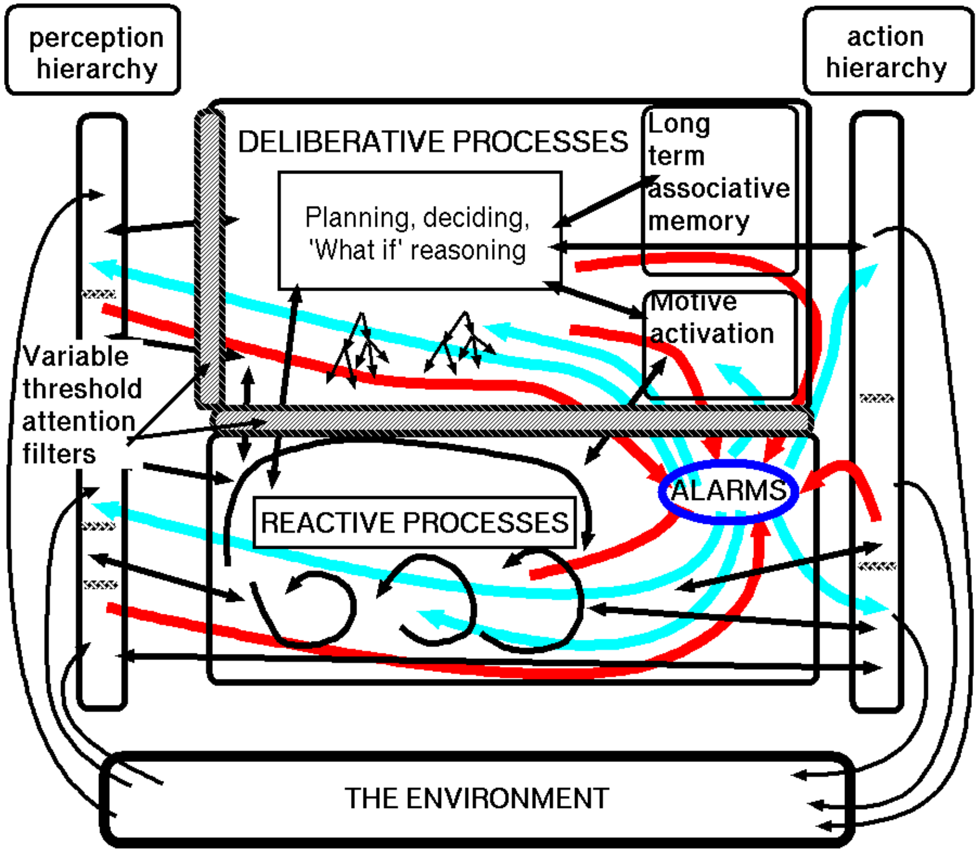
An alarm system can override, interrupt, abort, or modulate processing in other systems.

It can also make mistakes because it uses fast rather than careful decision making.

Primary and secondary emotions.

A deliberative system may need an alarm mechanism

With some additional mechanisms to act as attention filters, to help suppress some alarms and other disturbances during urgent and important tasks:



When powerful external factors override the attention filters this could amount to a certain sort of powerful emotional state.

Multi-window perception and action

If multiple levels and types of perceptual processing go on in parallel, we can talk about

“multi-window perception”,

as opposed to

“peephole” perception.

Likewise, in an architecture there can be

multi-window action

or merely

peephole action.

In multi-window perception, the sensory inputs can simultaneously feed different perceptual subsystems with different functions and different connections to more central mechanisms.

In multi-window action, multiple internal decisions and strategies for action, can simultaneously feed the motor subsystems, e.g. controlling not only which way you walk but how you walk (e.g. cautiously, slowly) and what else you do e.g. looking around for signs of an intruder....

Did Good Old Fashioned AI (GOFAI) fail?

It is often claimed that symbolic AI and the work on deliberative systems failed in the 1970s and 1980s and therefore a new approach to AI was needed.

New approaches (some defended by philosophers) included use of neural nets, use of reactive systems, use of subsumption architectures (Rodney Brooks), use of evolutionary methods (genetic algorithms, genetic programming) and use of dynamical systems (using equations borrowed from physics and control engineering).

The critics missed the point that many of the AI systems of the 1970s and 1980s were disappointing partly because they used very small and very slow computers (e.g. 1MByte was a huge amount of memory in 1980), partly because they did not have enough knowledge about the world, and partly because the architecture lacked self-monitoring capabilities: meta-management.

The emphasis on architectures helps us think more clearly about combining components required to match human capabilities.

Evolutionary pressure towards self-knowledge, self-evaluation and self-control

A deliberative system can easily get stuck in loops or repeat the same unsuccessful attempt to solve a sub-problem.

One way to prevent this is to have a parallel sub-system monitoring and evaluating the deliberative processes. If it detects something bad happening, then it may be able to interrupt and re-direct the processing.

(Compare Minsky on “B brains” and “C brains” in *Society of Mind*)

We call this meta-management. It seems to be rare in biological organisms and probably evolved very late.

As with deliberative and reactive mechanisms, there are many forms of meta-management.

Conjecture: the representational capabilities that evolved for dealing with self-categorisation can also be used for other-categorisation, and vice-versa. Perceptual mechanisms may have evolved recently to use these those representational capabilities in percepts.

Example: seeing someone else as happy, or angry.

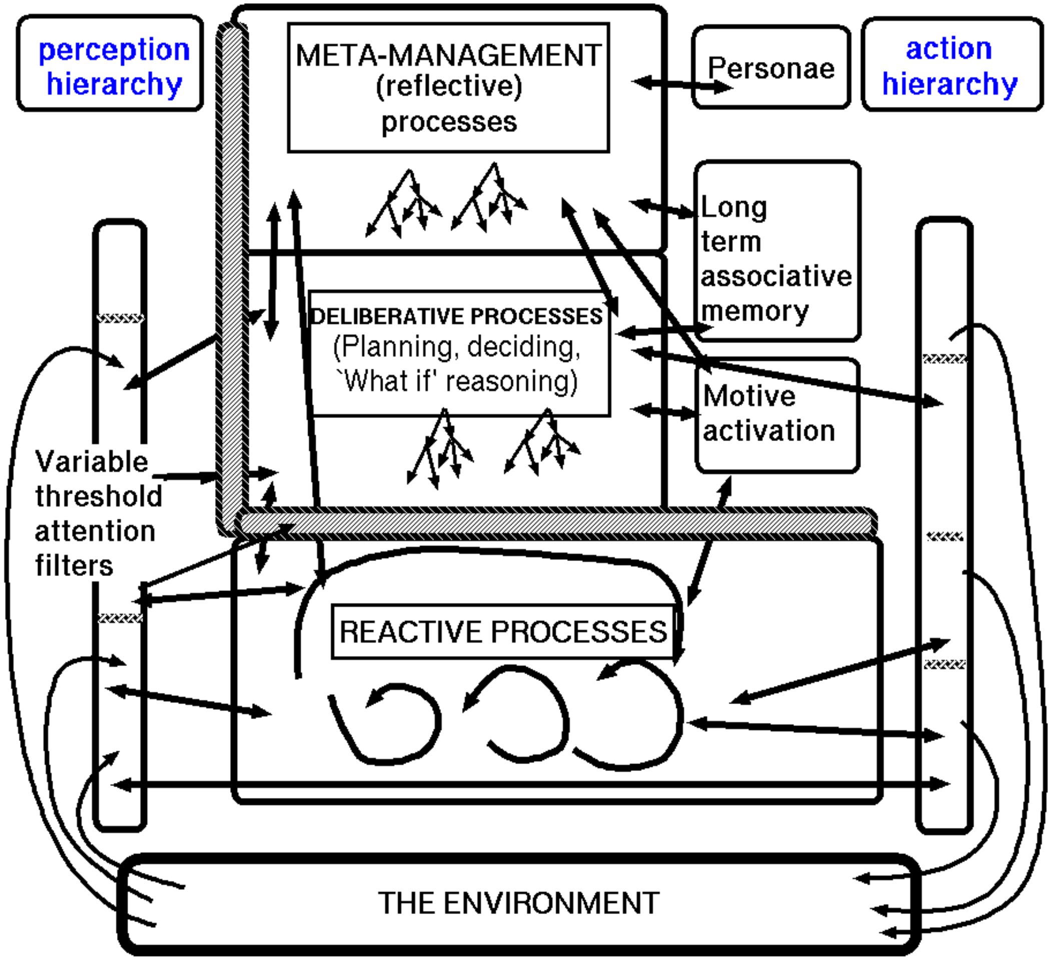
Later, meta-management (reflection) evolved

A conjectured generalisation of homeostasis.

Self monitoring, can include categorisation, evaluation, and (partial) control of internal processes.
Not just measurement.

The richest versions of this evolved very recently, and may be restricted to humans.

Research on ‘reflective’ AI systems is in progress.



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

Absence of meta-management can lead to stupid behaviour in AI systems, and in brain-damaged humans.

See A.Damasio (1994) *Descartes’ Error* (watch out for the fallacies).

Further steps to a human-like architecture

CONJECTURE:

Central meta-management led to opportunities for evolution of

- additional layers in ‘multi-window perceptual systems’

and

- additional layers in ‘multi-window action systems’,

Examples: social perception (seeing someone as sad or happy or puzzled), and stylised social action, e.g. courtly bows, social modulation of speech production.

Additional requirements led to further complexity in the architecture, e.g.

- ‘interrupt filters’ for resource-limited attention mechanisms,
- more or less global ‘alarm mechanisms’ for dealing with important and urgent problems and opportunities,
- socially influenced store of personalities/personae

All shown in the next slide, with extended layers of perception and action.

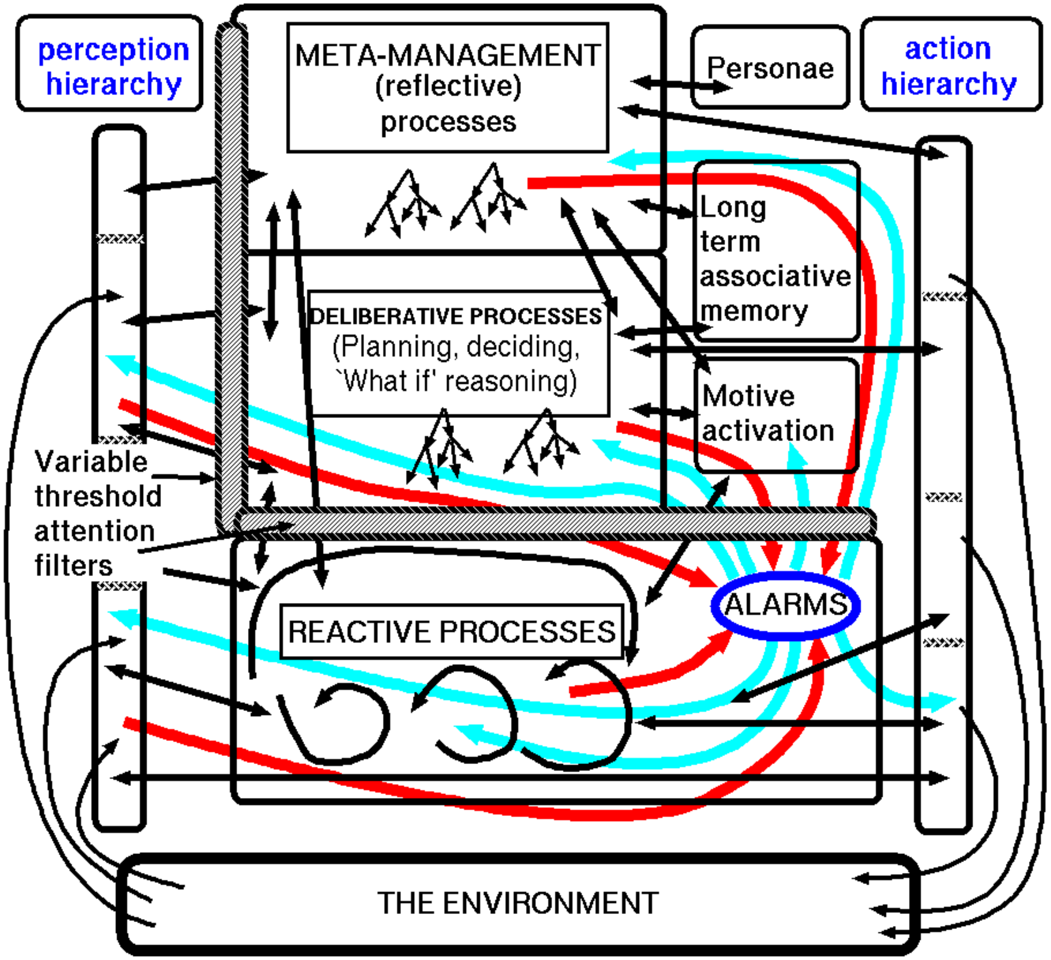
More layers of abstraction in perception and action, and global alarm mechanisms

This conjectured architecture (H-Cogaff) could be included in robots (in the distant future).

Arrows represent information flow (including control signals)

If meta-management processes have access to intermediate perceptual databases, then this can produce self- monitoring of sensory contents, leading robot philosophers with this architecture to discover “the problem(s) of Qualia?”

‘Alarm’ mechanisms can achieve rapid global re-organisation.


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

Meta-management systems need to use meta-semantic ontologies: they need the ability to refer to things that refer to things.

Some Implications

Within this framework we can explain (or predict) many phenomena, some of them part of everyday experience and some discovered by scientists:

- Several varieties of emotions: at least three distinct types related to the three layers: primary (exclusively reactive), secondary (partly deliberative) and tertiary emotions (including disruption of meta-management) – some shared with other animals, some unique to humans. (For more on this see Cogaff Project papers)
- Discovery of different visual pathways, since there are many routes for visual information to be used.
  (See this presentation

  `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk8`)
- Many possible types of brain damage and their effects, e.g. frontal-lobe damage interfering with meta-management (Damasio).
- Blindsight (damage to some meta-management access routes prevents self-knowledge about intact (reactive?) visual processes.)

This helps to enrich the analyses of concepts produced by philosophers sitting in their arm chairs: for it is very hard to dream up all these examples of kinds of architectures, states, processes if you merely use your own imagination.

Implications continued ....

- Many varieties of learning and development
  (E.g. "skill compilation" when repeated actions at deliberative levels train reactive systems to produce fast fluent actions, and action sequences. Needs spare capacity in reactive mechanisms, (e.g. the cerebellum?). We can also analyse development of the architecture in infancy, including development of personality as the architecture grows.)
- Conjecture: mathematical development depends on development of meta-management – the ability to attend to and reflect on thought processes and their structure, e.g. noticing features of your own counting operations, or features of your visual processes.
- Further work may help us understand some of the evolutionary trade-offs in developing these systems.
  (Deliberative and meta-management mechanisms can be very expensive, and require a food pyramid to support them.)
- Discovery by philosophers of sensory 'qualia'. We can see how philosophical thoughts (and confusions) about consciousness are inevitable in intelligent systems with partial self-knowledge.

See also the papers here: `http://www.cs.bham.ac.uk/research/cogaff/`

How to explain qualia

Philosophers (and others) contemplating the content of their own experience tend to conclude that there is a very special type of entity to which we have special access only from inside qualia (singular is ‘quale’). This generates apparently endless debates.

For more on this see talk 12 on consciousness here

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

We don’t explain qualia by saying what they are.

Instead we explain the phenomena that generate philosophical thinking of the sort found in discussions of qualia.

It is a consequence of having the ability to attend to aspects of internal information processing (internal self-awareness), and then trying to express the results of such attention.

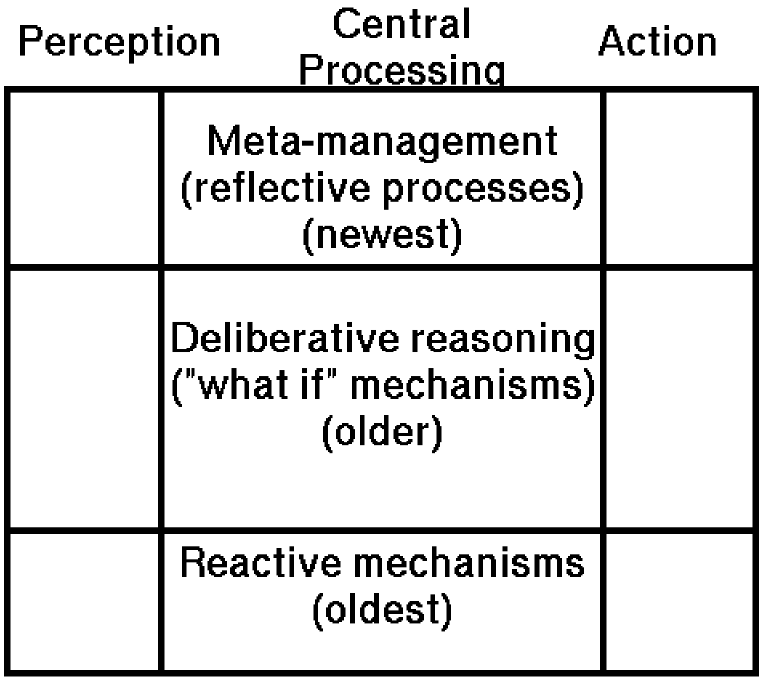
That possibility is inherent in any system that has the sort of architecture we call H-Cogaff, though different versions will be present in different architectures, e.g. depending on the forms of representation and modes of monitoring available to meta-management.

Robots with that architecture may also ‘discover’ qualia.

How to talk about architectures

Towards a taxonomy (ontology, perhaps a generative grammar), for architectures.

- **Types of information used**
- **Uses of information**
- **Forms of representation**
- **Types of Mechanism**
- **Ways of putting things together in an architecture**



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

Architectures vary according to which of the boxes contain mechanisms, what those mechanisms do, which mechanisms are connected to which others.

Also, architectures are not static: some contain contain the ability to grow and develop – new layers, new mechanisms, new forms of representation, new links between mechanisms – e.g. a new-born human's architecture.

NB The diagrams should indicate more clearly how sensory, motor, and more "central" subsystems can overlap in their mechanisms and functionality. Compare the diagrams in

`http://tinyurl.com/CogMisc/vm-functionalism.html`

Families of architecture-based mental concepts

For each architecture we can specify a family of concepts of types of virtual machine information processing states, processes and capabilities supported by the architecture.

Theories of the architecture of matter refined and extended our concepts of kinds of stuff (periodic table of elements, and varieties of chemical compounds) and of physical and chemical processes.

Likewise, architecture-based mental concepts can extend and refine our semantically indeterminate pre-theoretical concepts, leading to much clearer concepts related to the mechanisms that can produce different sorts of mental states and processes.

Philosophy will never be the same again.

Aristotle: The soul is the form of the body

21st Century: Souls are virtual machines implemented in bodies

Human souls are products of both evolution and development in a rich environment.

Artificial souls may be produced by designers or a mixture of evolutionary algorithms and learning in a rich environment, or ...?

New questions supplant old ones

We can expect to replace old unanswerable questions.

Is a fly conscious? Can a foetus feel pain?

is replaced by new EMPIRICAL questions, e.g.

Which of the 37 (or 370, or 3,700) varieties of consciousness does a fly have, if any?

Which types of pain can occur in an unborn foetus aged N months and in which sense of 'being aware' can it be aware of them, if any?

Of course, this may also generate new ethical questions, about the rights of robots and our duties towards them.

And that will feed new problems into moral philosophy.

The causation problem: Epiphenomenalism

A problem not discussed here is how it is possible for events in virtual machines to have causal powers.

It is sometimes argued that since (by hypothesis) virtual machines are fully implemented in physical machines, the only causes really operating are the physical ones.

This leads to the conclusion that virtual machines and their contents are "epiphenomenal", i.e. lacking causal powers.

If correct that would imply that if mental phenomena are all states, processes or events in virtual information processing machines, then mental phenomena (e.g. desires, decisions) have no causal powers.

A similar argument would refute many assumptions of everyday life, e.g. ignorance can cause poverty, poverty can cause crime, etc.

Dealing with this issue requires a deep analysis of the notion of 'cause', probably the hardest unsolved problem in philosophy.

A sketch of an answer is offered in this Philosophy of AI tutorial presentation:

`http://www.cs.bham.ac.uk/research/projects/cogaff/ijcai01`

See also the talk on supervenience and implementation in
smaller `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/`

A problem

Too many philosophers are ignorant of AI and Computer Science, and have over-simplified ideas about what they do.

So they produce false generalisations about what computers can or cannot do.

Too many AI researchers are ignorant of philosophy and do not have good philosophical analytical skills.

So they do not detect muddle and confusion in their own concepts.

MAYBE YOU CAN STUDY BOTH AND HELP TO IMPROVE BOTH DISCIPLINES IN THE FUTURE?

Similar comments can be made about psychology and AI, or neuroscience and AI.

Why scientists need to learn to think like engineers

Although AI has always had a strong scientific strand as well as the engineering strand, one of the important influences from the engineering to the science, is the emphasis on the need to understand not just mechanisms, but the various alternative functional requirements that mechanisms may need to satisfy.

Building systems to replicate observed behaviours can lead to shallow science if the functional requirements met by those behaviours are not understood.

Why do infants, toddlers, kittens and baby apes play in the way they do?

When a young child starts engaging in verbal interactions is the child

- trying to discover what language is talked by others,

  or

- trying to create a language that enables it to communicate effectively?

Why does the answer matter?

During the presentation I introduced questions in philosophy of mathematics and related them to questions about the evolution of cognitive systems.

Anyone interested in that may find this web site useful:

`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/mathsem.html`
From Molecules to Mathematicians:
How could evolution produce mathematicians from a cloud of cosmic dust?