Varieties of Consciousness

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

School of Computer Science
The University of Birmingham

Related papers can be found at

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

This and other slide presentations can be found at

http://www.cs.bham.ac.uk/˜axs/misc/talks/

With help from many students, friends and colleagues, including
Luc Beaudoin, Margaret Boden, Ron Chrisley, Stan Franklin, Catriona Kennedy, Brian Logan,
Matthias Scheutz, Ian Wright – and various luminaries from whom I've learnt.

Some questions: Let’s have a vote!

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?
- Is a new born baby conscious (when not asleep) ?
- Are you conscious when you are dreaming?
- Is the file protection system in an operating system conscious of attempts to violate access permissions?
- Is a soccer-playing robot conscious?

Advertisement

MICROSOFT-FREE ZONE

I USE ONLY

LINUX/UNIX SYSTEMS

AND FREE SOFTWARE

Including: Latex, dvips, ps2pdf

Diagrams are created using tgif, freely available from

http://bourbon.cs.umd.edu:8001/tgif/

ftp://ftp.cs.ucla.edu/pub/tgif/

Before asking what consciousness is, let's ask what an elephant is.

The Parable of the Blind Men and the Elephant
John Godfrey Saxe (1816-1887)

http://www.wvu.edu/~lawfac/jelkins/lp-2001/saxe.html

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

The First approached the Elephant
And, happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me, but the Elephant
Is very like a wall!"

The Second, feeling of the tusk,
Cried, “Ho! what have we here
So very round and smooth and sharp?
To me ‘tis very clear
This wonder of an Elephant
Is very like a spear!”

The Third approached the animal
And, happening to take
The squirming trunk within his hands,
Thus boldly up he spake:
"I see," quoth he, "The Elephant
Is very like a snake!"

The Fourth reached out an eager hand,
And felt about the knee:
"What most the wondrous beast is like
Is very plain," quoth he;
"Tis clear enough the Elephant
Is very like a tree!"

The Fifth, who chanced to touch the ear,
Said, “Even the blindest man
Can tell what this resembles most;
Deny the fact who can:
This marvel of an elephant
Is very like a fan!”

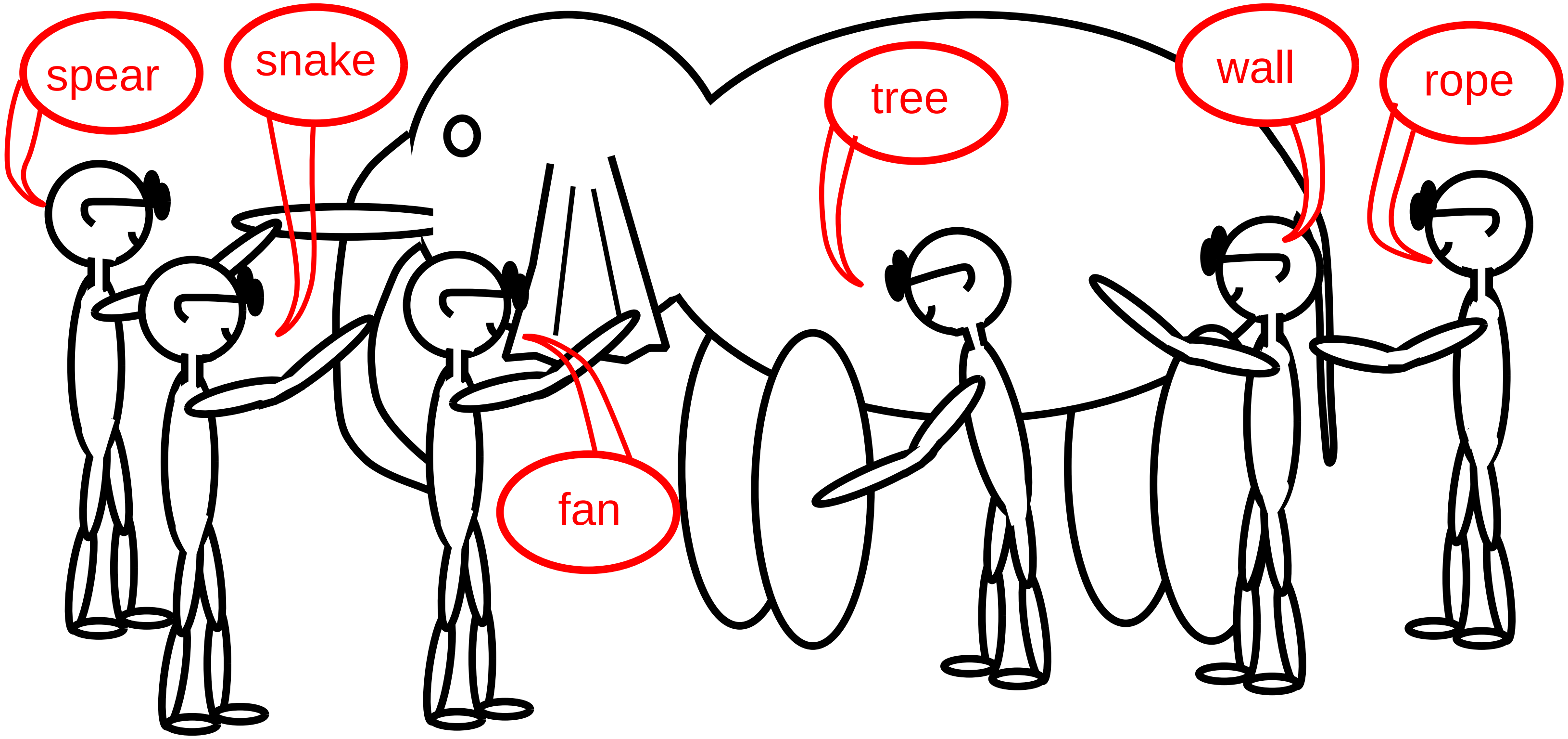
The Sixth no sooner had begun
About the beast to grope
Than, seizing on the swinging tail
That fell within his scope,
“I see,” quoth he, “the Elephant
Is very like a rope!”

And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong.
Though each was partly in the right,
They all were in the wrong!

(Use Google to search for “blind men elephant”)

What is an Elephant?

See: "The Parable of the Blind Men and the Elephant"
by John Godfrey Saxe (1816-1887)
http://www.wvu.edu/~lawfac/jelkins/lp-2001/saxe.html



Who can see the whole reality?

SUGGESTION

**Consciousness is
a huge elephant
studied by
many blind men
(and women)**

**(Actually several elephants,
as we'll see below.)**

Examples of what the blind gropers say and write about consciousness follow ....

Blind men describing consciousness

- It's indefinable, knowable only through having it
- It's what it is like to be something (hungry, in pain, happy, a bat...)
  (Compare http://www.cs.bham.ac.uk/~axs/misc/like_to_be_a_rock/)
- You lose consciousness when you are asleep
- You are conscious when you dream
- Consciousness is essential for processes to be mental
- Many mental processes are inaccessible to consciousness
- It causes human decisions and actions
- It has no causal powers (it is epiphenomenal)
- It can exist independently of physical matter (e.g. in an after-life)
- It's a special kind of stuff somehow produced by physical stuff
- It's just a collection of behavioural dispositions
- It's just a collection of brain states and processes

...continued

- It’s an aspect of a neutral reality which has both physical and mental aspects
- It’s just a myth invented by philosophers: best ignored
- It’s got something to do with talking to yourself (Dennett?)
- It’s something you either have or don’t have
- It’s just a matter of degree (of something or other)
- Consciousness requires a public (human) language
- Animals without language can have it
- All animals have it to some degree
- Humans are the only animals that have it
- It’s located in specific regions or processes in brains
- Talk about a location for consciousness is a “category mistake”

...continued

- Specific conscious events must have specific neural correlates
- Specific mental events are all multiply realisable, and therefore need not have fixed neural correlates.
- No machine could have it
- A machine that was indistinguishable from humans would have it
- Zombies are possible: machines that are indistinguishable from us could lack consciousness
- A machine that had exactly the same internal information processing capabilities as humans would necessarily have it.

.... and so on and so on

Exercise for students:

Find examples in philosophical and scientific literature of authors making those statements.

Can we do something about this babel?

Is there a whole elephant that nobody has yet seen all at once?

Or perhaps more than one elephant?

We need to find a way to step outside the narrow debating arenas to get a bigger picture.

We'll then see all the sub-pictures
at which myopic debaters peer,
and understand why their descriptions
are at best only part of the truth
and at worst just products of muddle, confusion,
ignorance and prejudice (e.g. religious prejudice).

Or value judgements: unwillingness to grant robots the right to have their desires taken seriously?

Revising the parable

And so these blind philosophers
(psychologists/brain scientists/...)
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong.
Though each was partly in the right,

They all were in the wrong!

Partial diagnosis (1)

Scientific and philosophical discussion of consciousness is

a real mess

for several reasons, including:

– **Much conceptual confusion**

(caused partly by unwitting use of ‘cluster concepts’
See: http://www.cs.bham.ac.uk/research/cogaff/sloman-lmpsfinal.pdf)

Partial diagnosis (2)

Scientific and philosophical discussion of consciousness is

a real mess

for several reasons, including:

– Much conceptual confusion
(caused partly by unwitting use of ‘cluster concepts’
See: http://www.cs.bham.ac.uk/research/cogaff/sloman-lmpsfinal.pdf)

– **Excessive focus on one case:**
normal adult (academic?) humans

Partial diagnosis (3)

Scientific and philosophical discussion of consciousness is

a real mess

for several reasons, including:

- Much conceptual confusion
  (caused partly by unwitting use of ‘cluster concepts’
  See: http://www.cs.bham.ac.uk/research/cogaff/sloman-lmpsfinal.pdf)
- Excessive focus on one case normal adult (academic?) humans
- **Limited ideas about possible types of machines**
  (due to deficiencies in our educational system)

Partial diagnosis(4)

Scientific and philosophical discussion of consciousness is
a real mess

for several reasons, including:

- Much conceptual confusion
  (caused partly by unwitting use of ‘cluster concepts’
  See: http://www.cs.bham.ac.uk/research/cogaff/sloman-lmpsfinal.pdf)
- Excessive focus on one case normal adult (academic?) humans
- Limited ideas about possible types of machines
  (due to deficiencies in our educational system)
- **Especially lack of understanding about virtual machines**
  (virtual information processing machines)
  (Even computer scientists do not all grasp the generality
  and importance of the idea, though they use it every day.)

Partial diagnosis (5)

Scientific and philosophical discussion of consciousness is

a real mess

for several reasons, including:

– **Much conceptual confusion**
(caused partly by unwitting use of ‘cluster concepts’
See: http://www.cs.bham.ac.uk/research/cogaff/sloman-lmpsfinal.pdf)

– **Excessive focus on one case normal adult (academic?) humans**

– **Limited ideas about possible types of machines**
(due to deficiencies in our educational system)

– **Especially lack of understanding about virtual machines**
(virtual information processing machines)

– **The illusion of “direct access” to the nature of consciousness**
(We experience it so directly there’s no room
for mistaken beliefs about it – or is there?)

Partial diagnosis (6)

Scientific and philosophical discussion of consciousness is

a real mess

for several reasons, including:

- **Much conceptual confusion (caused partly by unwitting use of 'cluster concepts'**
  See: http://www.cs.bham.ac.uk/research/cogaff/sloman-lmpsfinal.pdf)
- **Excessive focus on one case normal adult (academic?) humans**
- **Limited ideas about possible types of machines**
  (due to deficiencies in our educational system)
- **Especially lack of understanding about virtual machines**
  (virtual information processing machines)
- **The illusion of "direct access" to the nature of consciousness**
- **Blind men describing an elephant:**
  people ignore, or don't know, most of the relevant facts.

Summary diagnosis

– Much conceptual confusion
(caused partly by unwitting use of ‘cluster concepts’
See: http://www.cs.bham.ac.uk/research/cogaff/sloman-lmpsfinal.pdf)

– Excessive focus on one case normal adult (academic?) humans

– Limited ideas about possible types of machines
(due to deficiencies in our educational system)

– Especially lack of understanding about virtual machines
(virtual information processing machines)

– The illusion of “direct access” to the nature of consciousness

– Blind men describing an elephant:
people ignore, or don’t know, most of the relevant facts.

A golden rule for studying consciousness:

DO NOT ASSUME THAT YOU CAN GRASP THE NATURE OF CONSCIOUSNESS. SIMPLY BY LOOKING INSIDE YOURSELF, HOWEVER LONG, HOWEVER CAREFULLY, HOWEVER ANALYTICALLY,

- **Introspection is merely one of many types of perception.**
- **Like other forms of perception it provides only information that the perceptual mechanism is able to provide. (That's a tautology!)**
- **Compare staring carefully at trees, rocks, clouds, stars, birds and beasts hoping to discover the nature of matter.**
- **At best you learn a subset of what needs to be explained, like perceiving only the elephant's trunk.**

Instead of gazing at our internal navels

We need to collect far more data-points, e.g. concerning:

- differences between humans at various stages of development,
- differences between mental phenomena in different cultures,
- unobvious aspects of conscious experiences,
  (including unconscious aspects)
- surprising effects of brain damage or disease,
- similarities and differences between different species,
- stages and trends in evolution, ...

And more importantly

We need deeper, richer forms of explanatory theories

– able to accommodate ALL the data-points

– which are mostly qualitative, not quantitative,

– and are mostly concerned with what can happen or can be done, rather than with laws or correlations.

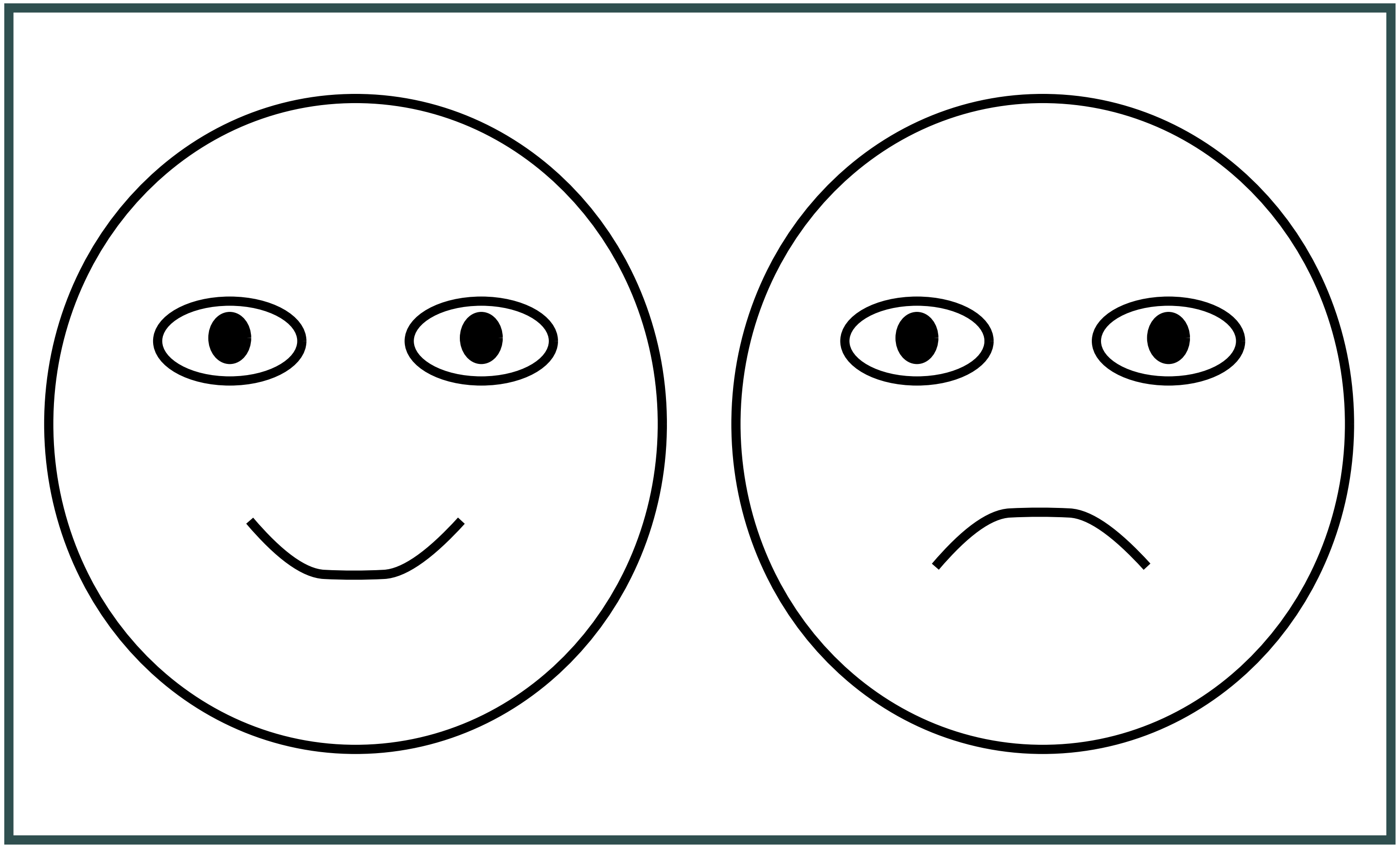
Note that the language of physics (mainly equations) is not as well suited to describing these realms of possibility as the languages of formal linguistics (grammars of various kinds) and the languages of computer scientists, software engineers and AI theorists (including languages which specify machines that interpret new languages).

The latter are languages for specifying and explaining the behaviour of information processing machines.

We may find that no existing languages are adequate for our purposes.

Of course we can use introspection

E.g. what's the difference between how these two faces look?



Do the eyes look different? How? (Not everyone sees the difference.)
For those who do, describing it is quite hard.

How can you see eyes as happy or sad?

For someone who sees eyes as happy or sad it may feel like an unanalysable experience, yet, such seeing must involve use of the concepts “happy” and “sad”.

What is involved in having such concepts? Or applying them?

How are the concepts applied to a pair of black elliptical patterns on a white background?

How can the application of the concept depend not just on what is in the pattern (since the two pairs of eyes are physically identical) but on the context?

When those concepts are applied to those patterns, there must be a lot going on of which we are completely unaware.

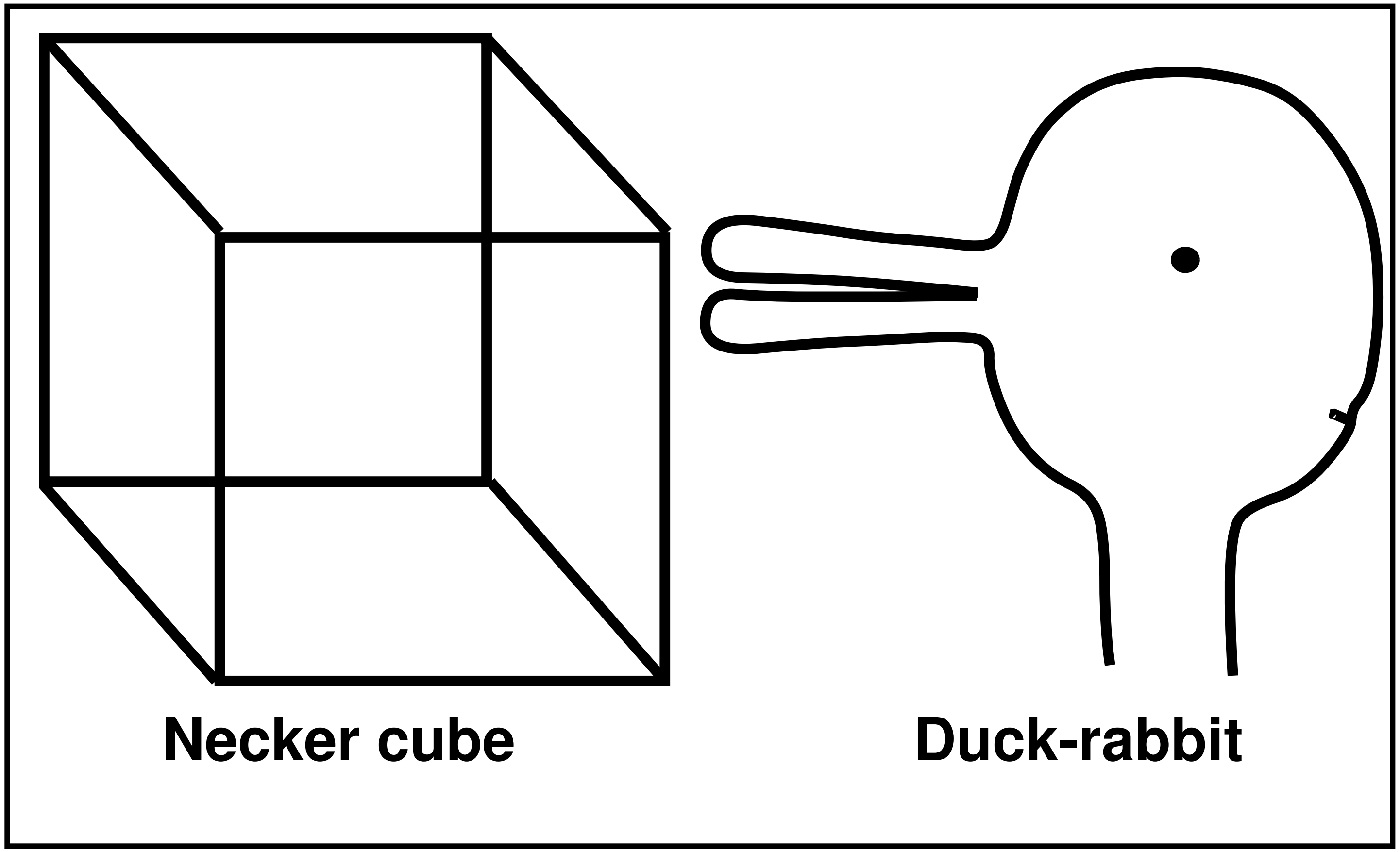
Conjecture: consciousness as we know it is necessarily but the tip of an iceberg of information-processing that is mostly totally inaccessible to consciousness.

(Compare the discussion of ‘Seeing as’ in

L.Wittgenstein (1953), *Philosophical Investigations*, Part 2, section xi.)

Another example

Both figures are ambiguous, and can 'flip'



What exactly changes when they flip?
One change is purely geometrical, one far more subtle.

Necker Cube and Duck-rabbit

When the Necker cube figure flips, all the changes are geometric.

They can be described in terms of relative distance and orientation of edges, faces and vertices.

When the duck-rabbit flips the geometry does not change:

- The functional interpretation of the parts changes (“bill”, “ears”).
- More subtle features change, attributable only to animate entities. E.g. “Looking left”, or “looking right”.

What does it mean to say that you “see the rabbit facing to the right”.

Perhaps it involves seeing the rabbit as a potential mover, more likely to move right than left.

Or seeing it as a potential perceiver, gaining information from the right.

What does categorising another animal as a perceiver involve? How does it differ from categorising something as having a certain shape?

Is this seeing, or only inferring?

The differences between different experiences

of the same ambiguous picture are visual, not simply inferential.

The examples occur in textbooks on *vision*, not *reasoning*.

Seeing mental states

**What is involved in seeing an "expression"
e.g. happiness, sadness?**

It is NOT just a matter of recognising and labelling a pattern.

Those visual categories are semantically linked to matters of importance to us as social animals,

**just as the perception of geometric structure
is linked to our needs as agents in complex 3-D world
and our ability to act in that world.**

Seeing how someone feels can affect what you should do next: a non-geometric kind of affordance. and it seems to 'colour' the whole percept.

An appropriate architecture should explain the ability to have this sort of percept. (See H-Cogaff, later.)

For more on visual affordances see talk 8 here:
http://www.cs.bham.ac.uk/˜axs/misc/talks/

So the experience of seeing has “hidden richness”

What constitutes our experience at any time is

a large collection of unrealised, unactivated,

but potentially activatable capabilities,

in addition to a large collection that we unaware of activating.

Can we say more about what those unacknowledged capabilities are?

One way is to learn from psychologists and brain scientists about the many bizarre ways they can go wrong.

We can also learn new ways of looking at old experiences.

E.g. how exactly do you experience an empty space?

(Most AI vision systems can’t)

Humans (e.g., painters, creators of animated cartoons, etc.) can experience an empty space as full of possibilities for shapes, colours, textures, and processes in which they change. How?

Can any other animal? What sort of machine could?

So introspective analysis of experience can be useful

But doing it well can be very difficult, since most of what makes an experience what it is is not part of the experience.

In part what makes something the experience it is, is the collection of possibilities for change inherent in that experience – a sort of grammar.

And all those possibilities permit a range of interpretations.

Beyond introspection

So there's far more besides what is actually experienced that needs to be understood and explained.

In particular, what sort of machinery makes it possible?

Insofar as different organisms, or children, or people with various sorts of brain damage or disease, have different kinds of mental machinery, the types of experiences possible for them will be different.

Understanding our own case involves seeing how we fit into the total picture of biological evolution and its products.

Including other possible systems on other planets, and also in future robot labs.

There's many more elephants than meet the eye – wherever you look.

So let's start again

BASIC WORKING ASSUMPTION: No magic, only complex engines

- Consciousness started as a biological phenomenon
- It was produced by evolution, somehow using physical resources
- But that does not make it a physical phenomenon (since lots of non-physical things are produced using physical resources, e.g. poverty, legal obligations.).
- It is not one thing: there are many varieties of consciousness in biological organisms - we need a new conceptual framework for thinking about how they differ and how they are similar.
- They all depend on the fact that biological organisms are information processors
- We can abstract away from some of the specifics of evolved life on earth to explore more varieties of information processors, as long as we take care to learn from the biological subset.

So let's start again... continued

- Suitable non-biological machines should, in principle, be able to replicate most aspects of biological forms of consciousness.
- Future human-like machines will re-discover all the puzzles of consciousness that have befuddled humans:

  because they have the same or very similar mental features.
  (NB: not all humans are alike).

REMEMBER: There are at least three types of machines

- Matter manipulating machines
  Diggers, drills, cranes, cookers
- Energy manipulating machines
  Diggers, drills, cranes, cookers, transformers, steam engines...
- Information manipulating machines
  Thermostats, Controllers, most organisms, operating systems, compilers, ...

The engines, the mechanisms that makes information manipulation possible, are not just physical machines, e.g. made of blood, meat, etc. They include also virtual machines.

Virtual vs physical machines

In computer science, software engineering and AI we have learnt the importance of virtual machines, e.g. the Lisp, Prolog, Java virtual machines, chess virtual machines, spreadsheets, neural nets, etc.

Virtual machines are *implemented* in physical machines.

Mechanisms that operate on complex information structures are typically virtual machines rather than physical machines. E.g.

- parsers,
- compilers,
- structure matchers,
- search engines,
- planners, ... etc.

If we are to explore the full range of designs for behaving systems, we need to be familiar with a wide range of techniques for constructing virtual machines of various sorts.

This has implications for the sorts of education that should be provided broad-minded students of brain, mind and behaviour (natural or artificial).

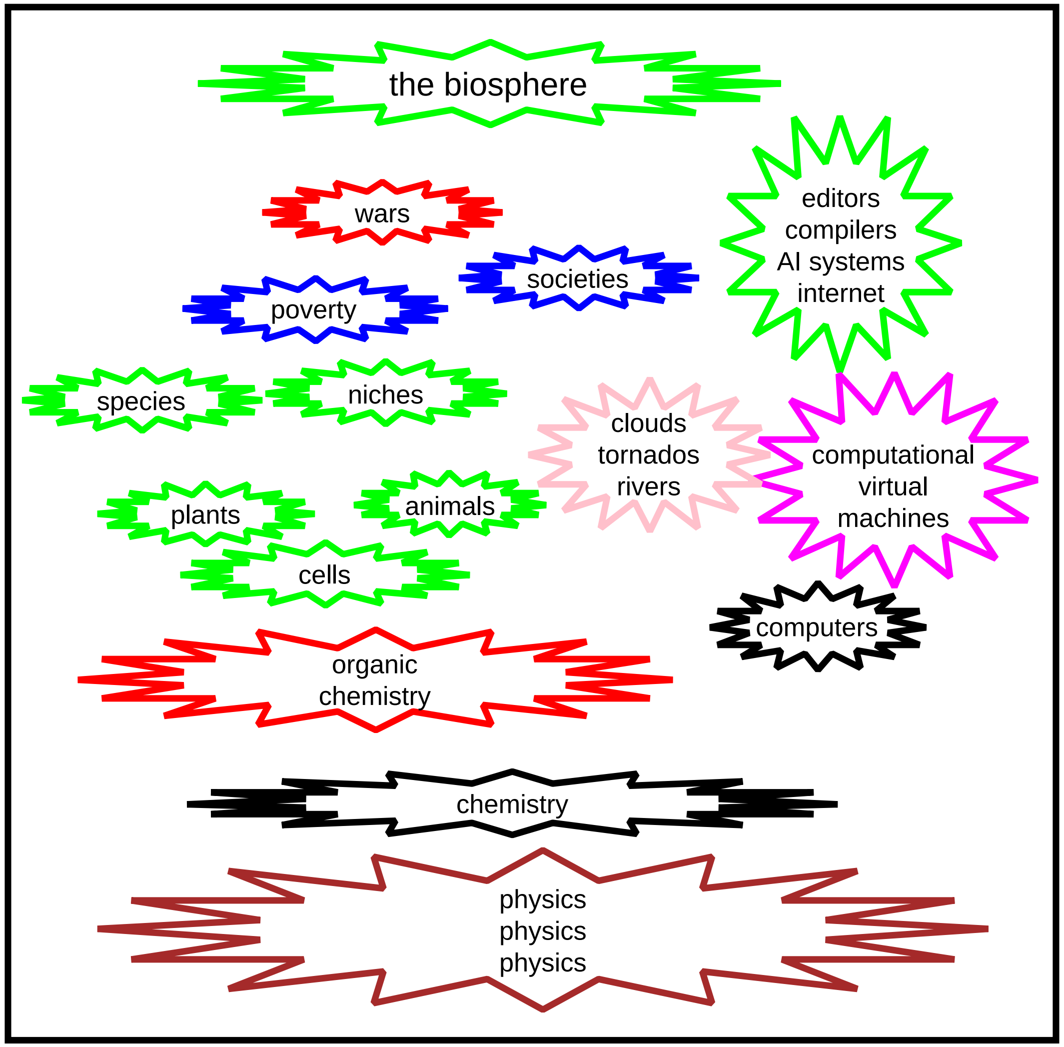
How to think about non-physical levels in reality

At all levels there are objects, properties, relations, structures, mechanisms, states, events, processes and causal interactions.

E.g. poverty can cause crime.

But they are all ultimately implemented in physical systems.

Nobody knows how many levels of virtual machines physicists will eventually discover.



See our IJCAI'01 Philosophy of AI tutorial http://www.cs.bham.ac.uk/~axs/ijcai01/

Evolution of information processing virtual machines

Evolution “discovered” and used many things long before human engineers and scientists thought of them.

Paleontology shows the development of physiology and provides some weak evidence about behavioural capabilities.

But there is very little direct evidence regarding previous forms of information processing: virtual machines leave no fossils.

Forms of information processing now found in nature give clues, and we can test theories in working models.

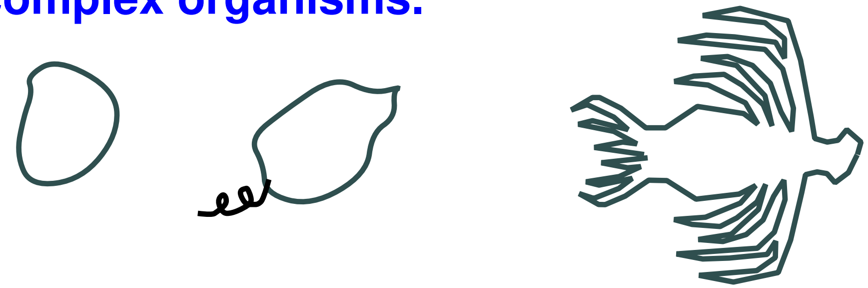
Some of the forms are evolutionarily very old. Others relatively new. (E.g. the ability to learn to read, design machinery, do mathematics, or think about your thought processes.)

WE NEED TO LEARN HOW TO ASK NEW (DEEP) QUESTIONS ABOUT THE POWERS AND FUNCTIONS OF DIFFERENT SYSTEMS, AND HOW THEY FIT TOGETHER.

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. But more interesting was their ability to *initiate* action, and to *select* responses, instead of simply being pushed around by resultants.

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

We use “information” in the everyday sense, not the Shannon/Weaver technical sense

Acting or selecting requires information

E.g. information about

- density gradients of nutrients in the primaeval soup
- the presence of noxious entities
- where the gap is in a barrier
- precise locations of branches in a tree as you fly through
- how much of your nest you have built so far
- which part of the nest should be extended next
- where a potential mate is
- something that might eat you
- something you might eat
- what that thing over there is likely to do next
- how to achieve or avoid various states
- how you thought about that last problem
- whether your thinking is making progress

and much, much more... (has anyone attempted a taxonomy?)

Most of these processes don't involve self-consciousness.

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.

Things you can do with information

A partial analysis to illustrate the above:

- You can react immediately (it 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 try to derive new information from it (e.g. what caused this? what else is there? what might happen next? can I benefit from this?)
- You can store it 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

... 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

An architecture includes

– forms of representation,

– algorithms,

– concurrently processing sub-systems,

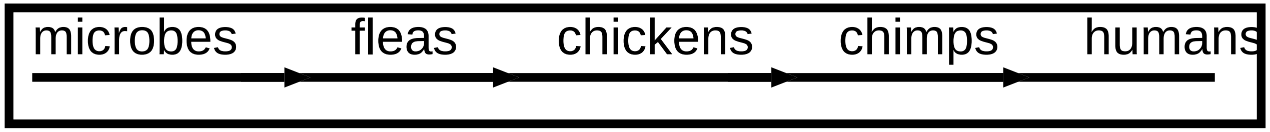
– connections between them.

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

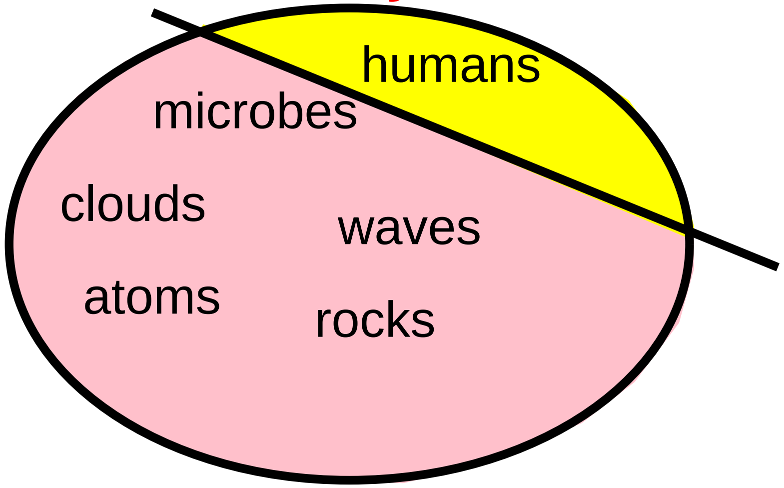
There's No Unique Correct Architecture

Some tempting wrong ways to think about consciousness:

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



2. It's not a dichotomy either



Intentionality and semantics

Intentionality involves the ability to refer to something (in thoughts, desires, plans, explanations, questions, etc.) I.e. it involves semantics (meaning).

Often assumed to be a requirement for consciousness: consciousness is always OF something.

John Haugeland distinguished derivative and original intentionality. Printed words, maps, footprints, etc. have derivative intentionality: they can only be used by something else to refer.

Philosophers often write as if “original intentionality” is either present or absent.

By exploring the variety of architectures for machines that process information we can distinguish a wide range of cases: varieties of intentionality.

Likewise if we look at many natural phenomena, e.g. people with brain damage, newborn infants, different sorts of animals.

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.).

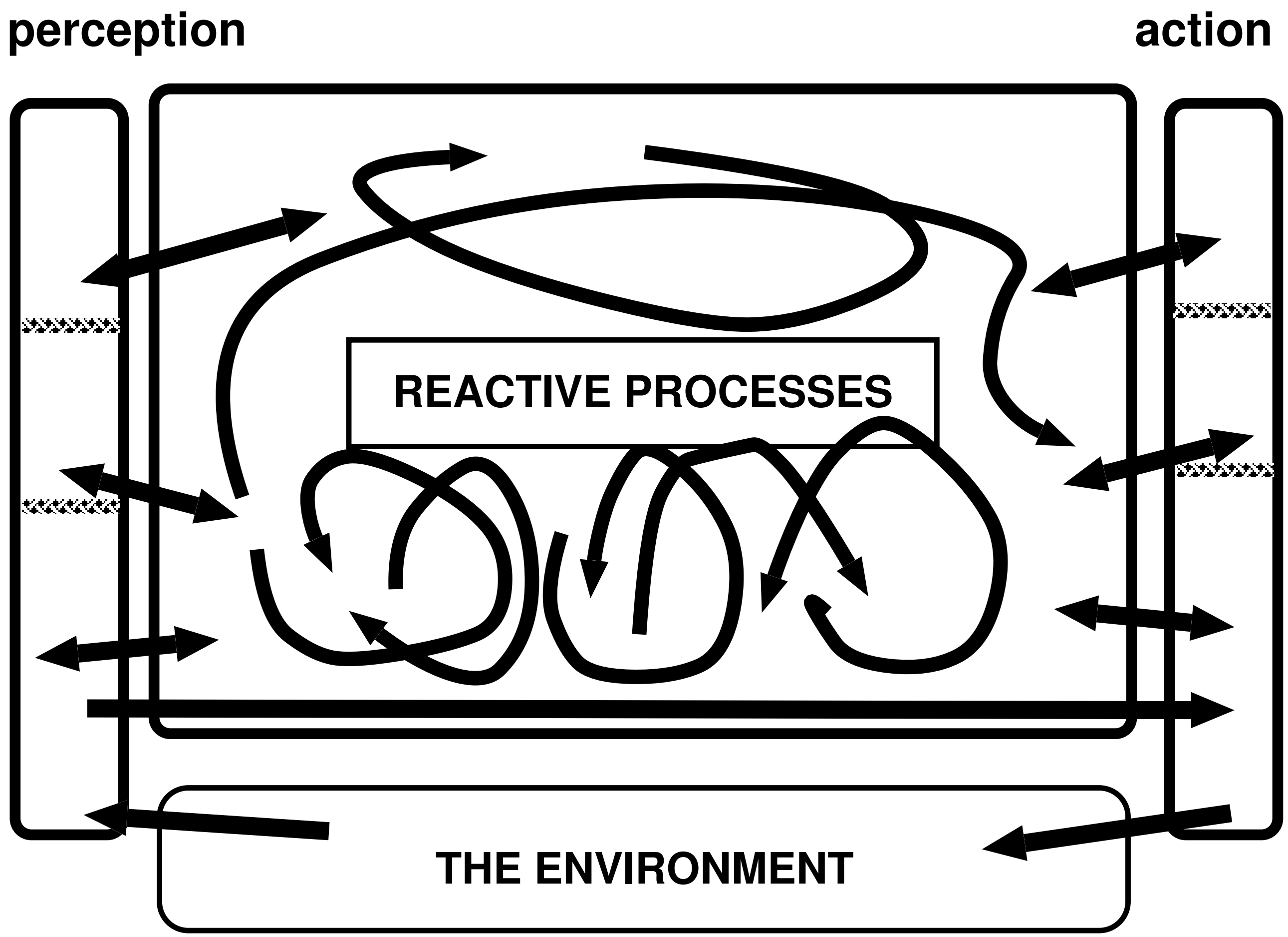
Some variations are discontinuous:

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

Most biological changes are discontinuous — discontinuities can be big or small.

In particular, evolution produces changes of kind as well as degree.

A simple (insect-like) architecture



An adaptive system with reactive mechanisms can be a very successful biological machine. Some purely reactive species also have a social architecture.

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 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.

“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 ways of using 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.

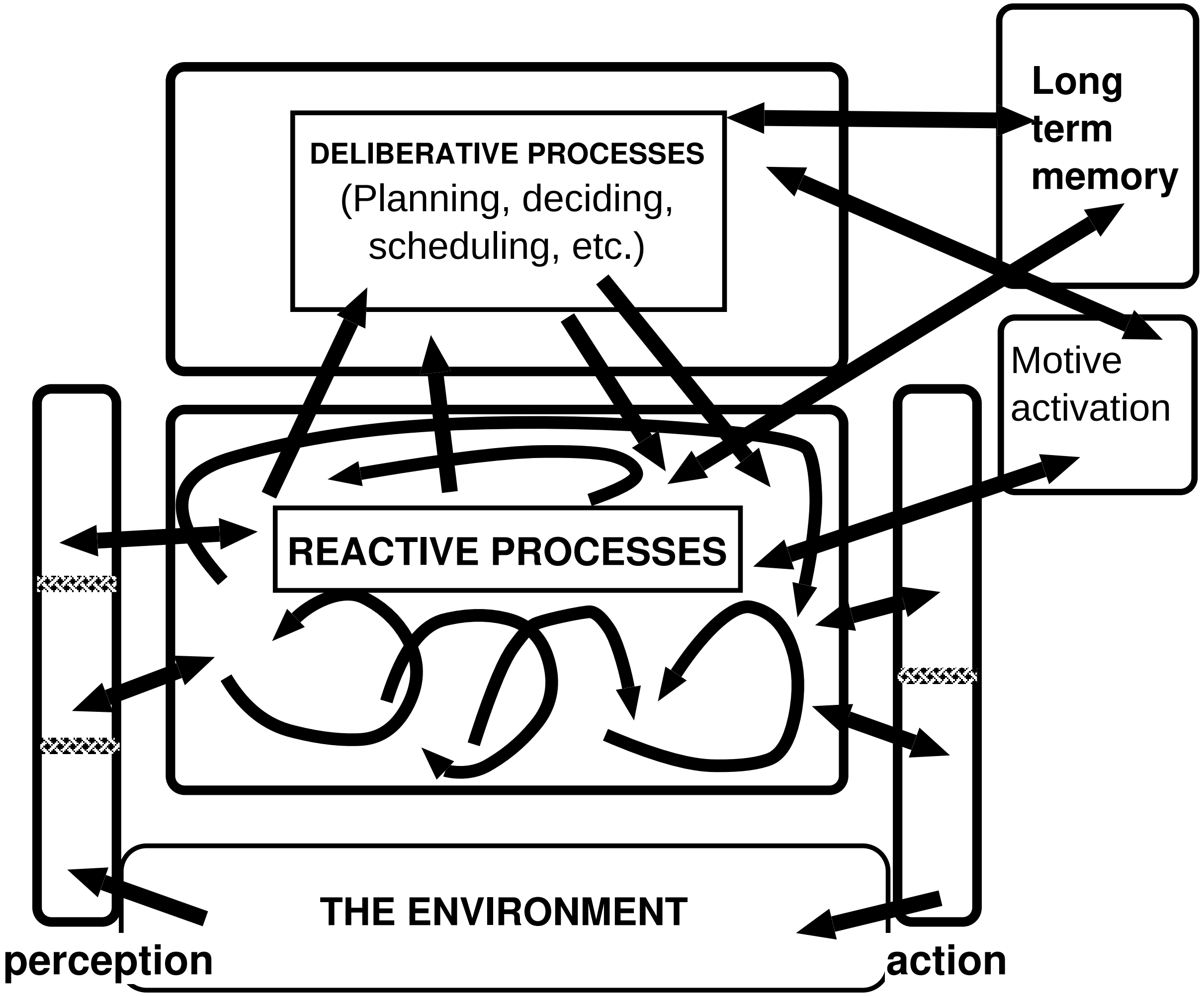
They also lack self-awareness, self-categorising abilities. A fly that sees an approaching hand probably does not know that it sees — it lacks meta-management mechanisms, described later.

Give REACTIVE DEMO

Sheepdog

‘Emotional’ agents

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).

Give DELIBERATIVE DEMO

SHRDLU (pop11 gblocks)

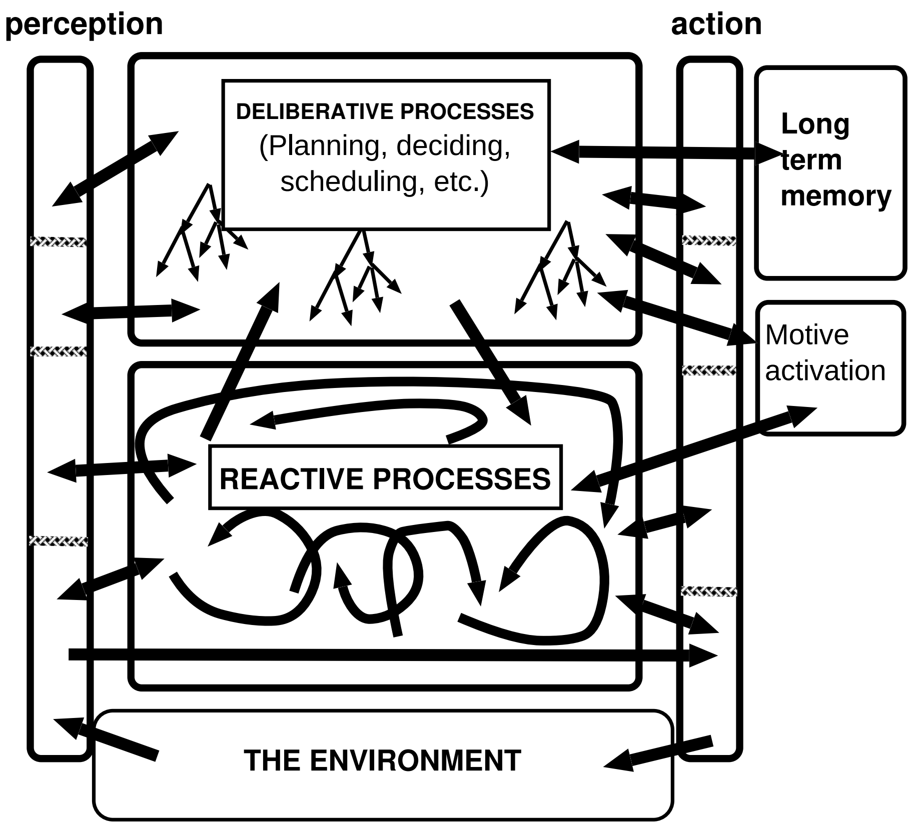
Deliberative mechanisms

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

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.

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.

The 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

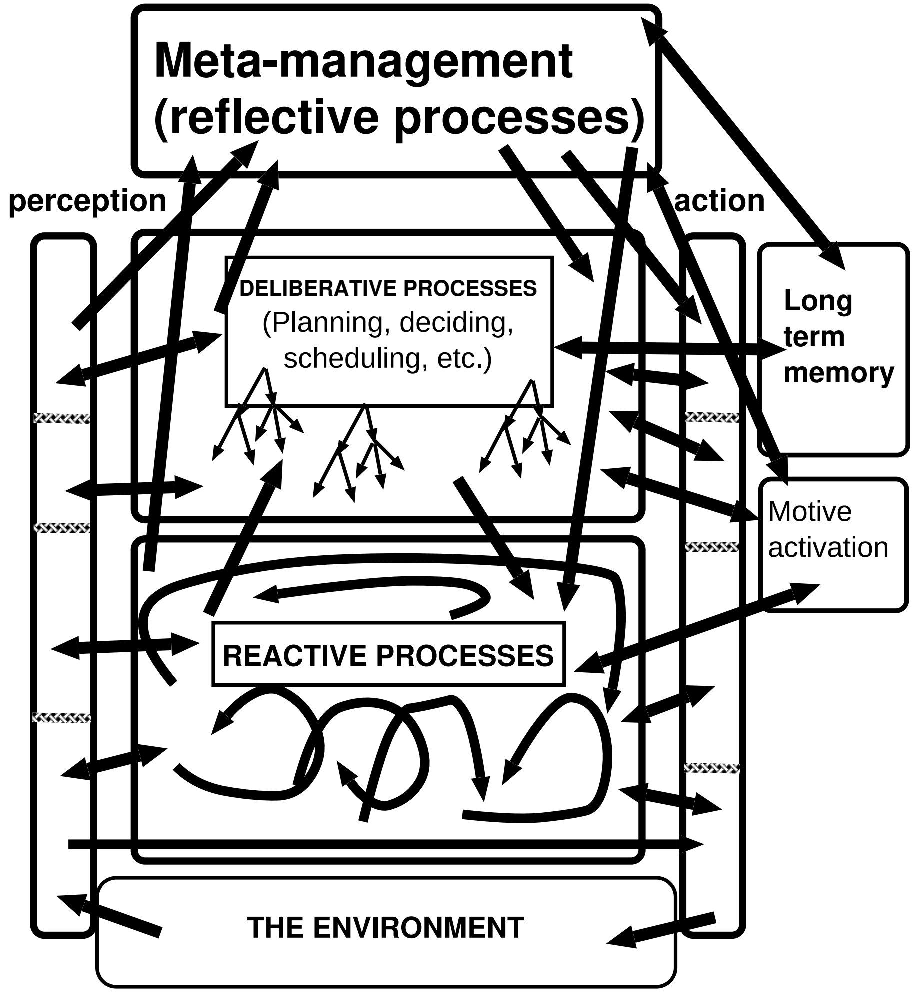
A conjectured generalisation of homeostasis.

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.

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



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

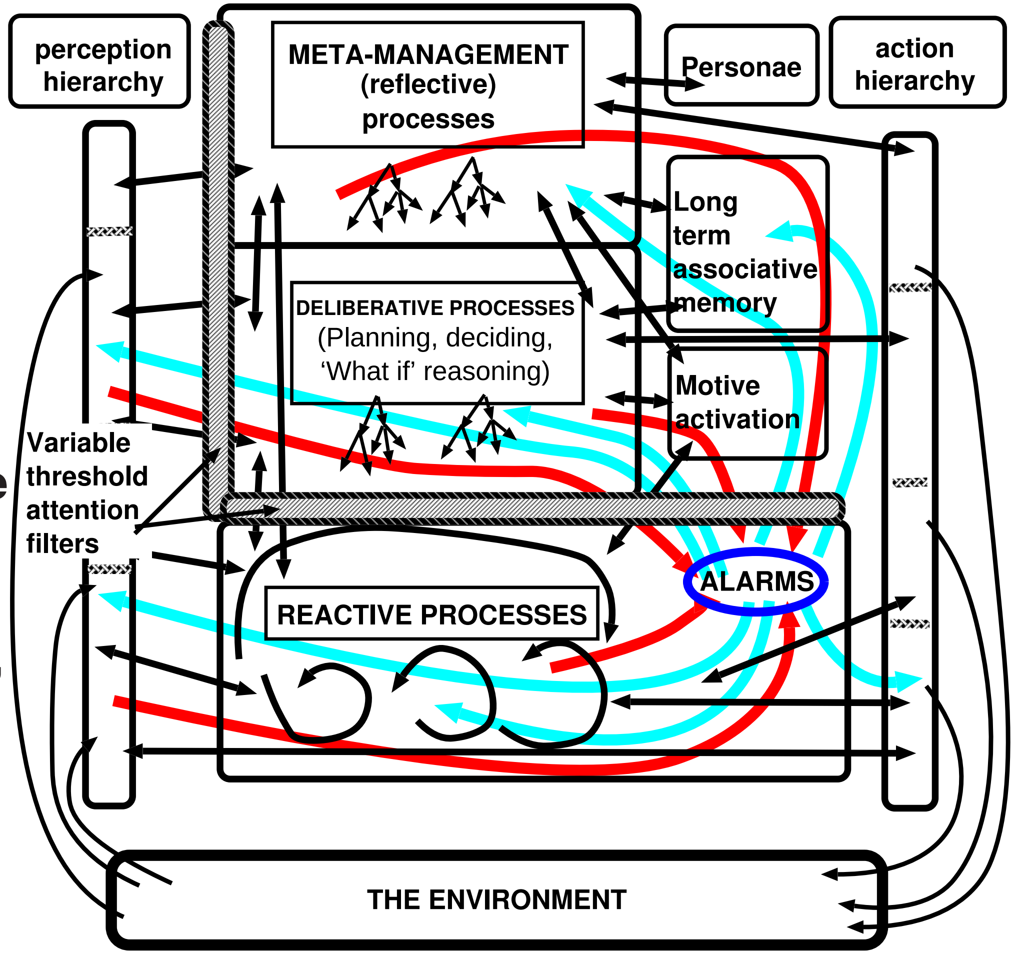
All shown in the next slide, with extended layers of perception and action.

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?"



For more detailed explanations of the ideas see
http://www.cs.bham.ac.uk/research/cogaff/0-INDEX00-02.html#74

Some Implications

Within this framework we can explain (or predict) many phenomena, some 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 talk 8 in http://www.cs.bham.ac.uk/~axs/misc/talks/)
- 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.)

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.**

For more see papers here: http://www.cs.bham.ac.uk/research/cogaff/

How to explain qualia

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

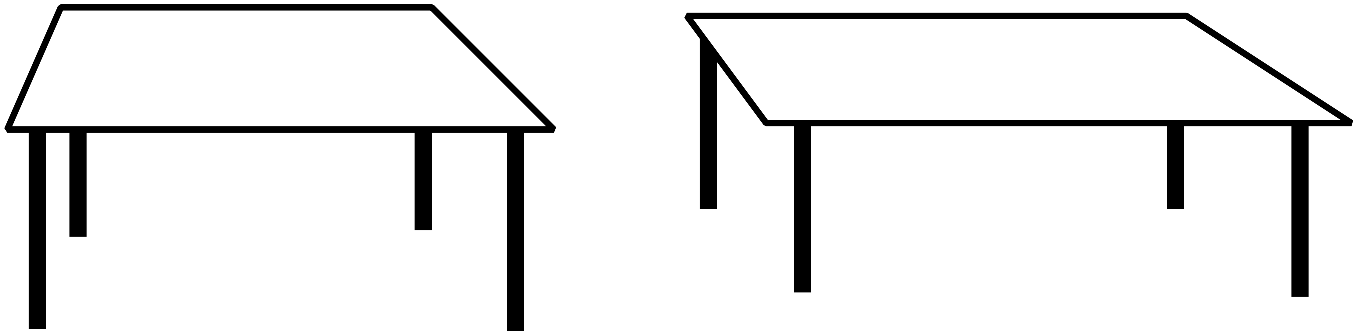
Instead we explain the phenomena that generate philosophical thinking of the sort found in discussions of qualia.

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.

And then ....

Table qualia



Six reflective robots

discussing their experience of the same table

seen from different viewpoints

could get bogged down

discussing consciousness

like six blind philosophers...

A new kind of explanation?

Notice that we are not saying: This is what qualia are .... Instead, we offer (conjectured) *sufficient* conditions for an information processing system to go through the very same processes as led humans to start thinking about sensory qualia and also other kinds. It's a side-effect of sophisticated biological mechanisms.

A meta-management system could give an agent the ability to attend not only to what is perceived in the environment, but to also features of the *mode of perception* that are closely related to properties of intermediate sensory data-structures.

Thus you can attend not only to (a) the table and its fixed 3-D shape, but also to (b) the 2-D *appearance* of the table in which angles and relative lengths of lines change as you change your viewpoint (or the table is rotated). The appearance can also change as you squint, tap your eyeball, put on coloured spectacles, ...

This is exactly the sort of thing that led philosophers (and others) to think about qualia as something internal, non-physical, knowable only from inside, etc.

It suffices to explain the mental mechanisms that generate an interest in thinking and talking about qualia. Robots with our information processing architecture would do the same.

Multiple elephants

The multi-disciplinary view of the whole architecture, and the different capabilities, states, processes, causal interactions, made possible by the various components, presents a (fairly) complete elephant.

(At least more complete than normal).

But there are different architectures, with very different information processing capabilities, supporting different states and processes.

E.g. fleas, fishes, philosophical humans.

So there are many elephants – not just one.

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.

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 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?

Biological changes are mostly discontinuous

The view that consciousness is just a matter of degree ignores the fact that evolutionary and developmental changes in biology are inherently discontinuous and involve many changes of kind.
E.g. structural changes, and development of new capabilities.

Why? (a) because molecular structures are discrete, (b) there can be only a finite number of generations between any two time points, which rues out a continuum of stages.

Likewise niches, and sets of requirements, can change discontinuously, depending on how surrounding designs change.

Some discontinuities may be big, others small: so not every space is either a continuum or a dichotomy.

All of this is compatible with Darwinian evolution.

A full analysis would explore trajectories in design space and niche space.

For more on this see talk 6 here: http://www.cs.bham.ac.uk/˜axs/misc/talks/

Mechanisms need an architecture

Different biological information processing functions require different kinds of mechanisms, often operating on different forms of representation and different forms of long and short term storage.

Sometimes they require different sub-mechanisms working together (perceiving, learning, using prior knowledge, deciding what to do, constructing plans, executing plans, etc.)

But there must always be an ARCHITECTURE combining all the mechanisms and processes they produce.

Some of the more sophisticated mechanisms and architectures evolved only relatively recently, and are in very few species (e.g. deliberative capabilities.)

We need to understand how they differ from, how they are built on, and how they interact with the much older, more wide-spread mechanisms.

The same organism, e.g. a human being, may include both very ancient commonplace mechanisms and very new rare mechanisms, in many sub-systems.

Evolution, the great philosopher/designer

In particular,

Evolution solved the "other minds problem" before anyone formulated it, by providing built-in apparatus for conceptualising mental states in others:

A requirement for

– prey species,
– predator species,
– social species.

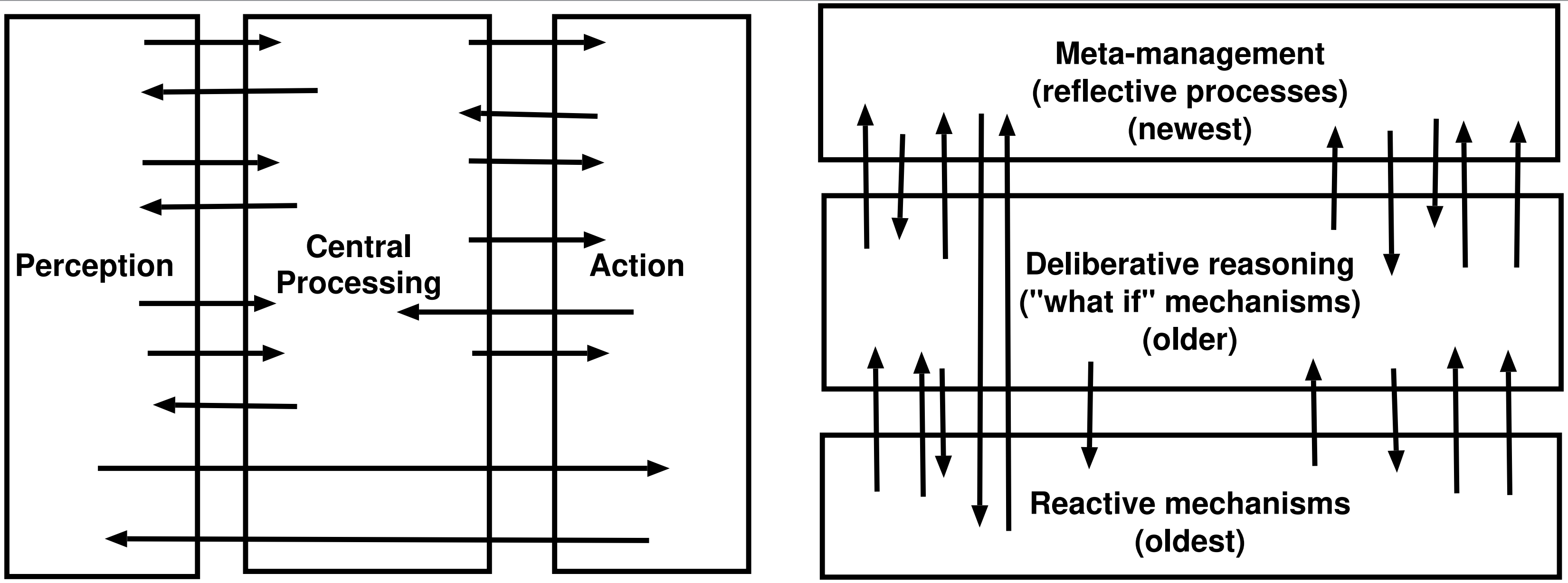
We need an architectural framework in which to place all these diverse capabilities, as part of the design task.

Later we can modify the framework as we discover its limitations.

The framework should simultaneously help us understand the evolutionary process and the results of evolution.

We have proposed the CogAff schema as a framework for thinking about a wide variety of information processing architectures, including both naturally occurring and artificial ones.

Towards an architecture schema



Two coarse divisions within information processing architectures – ‘towers’ and ‘layers’:

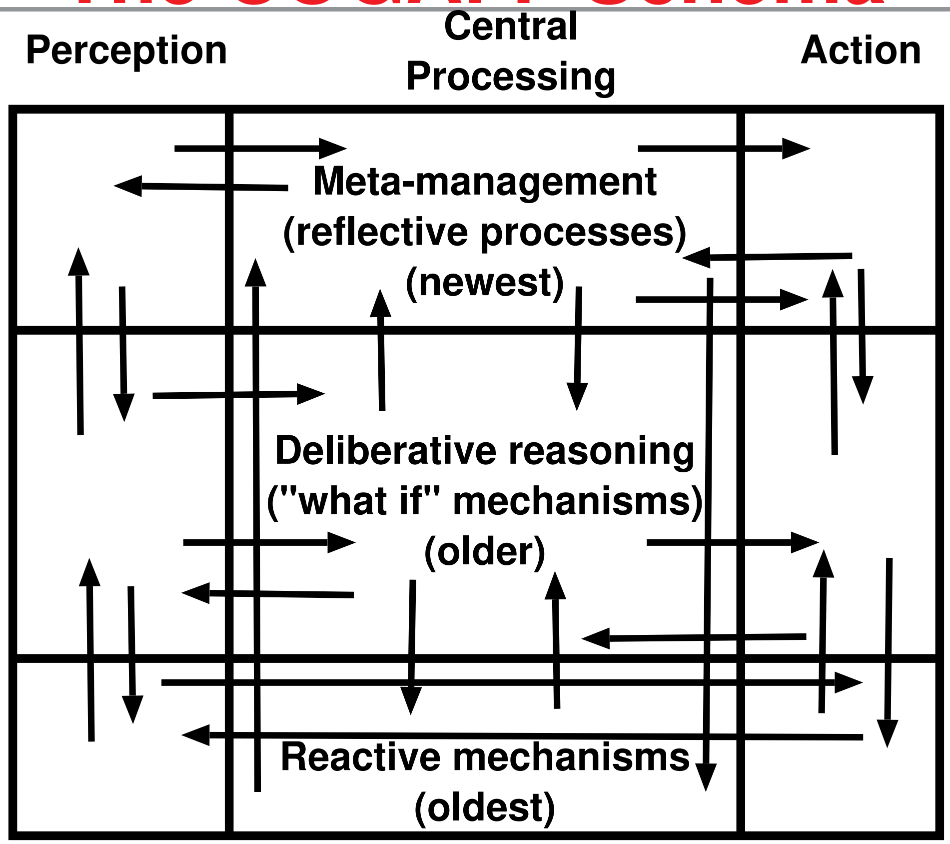
(a) Nilsson’s (1998) “triple tower” model

(b) Layered architectures: e.g. reactive, deliberative and meta-management layers.

(a) and (b) express different (orthogonal) functional divisions.

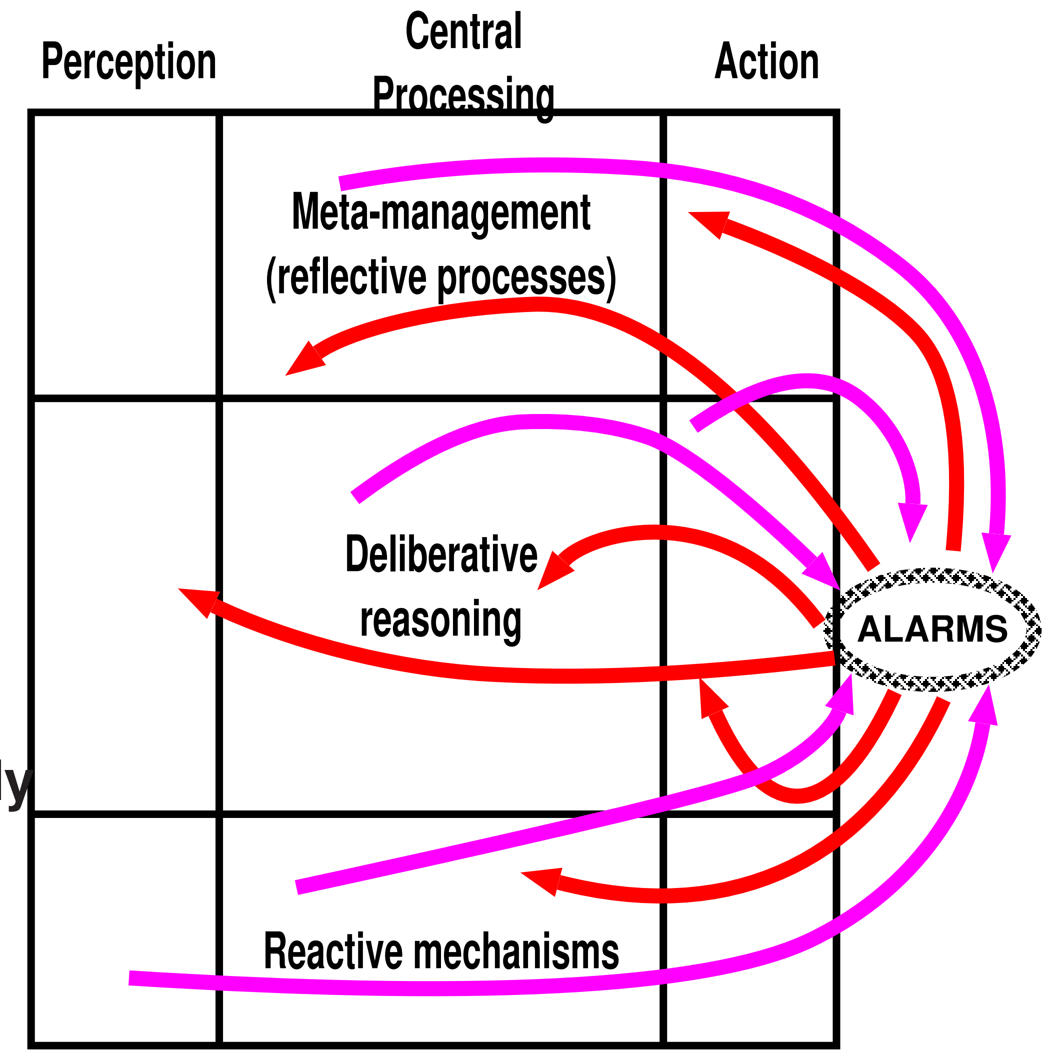
These divisions can be combined, as follows ....

Superimposing the divisions: The COGAFF Schema



Boxes indicate possible functional roles for mechanisms: only some possible information flow routes are shown (cycles are possible within boxes, but not shown).

COGAFF extended – with "alarm mechanisms"



Alarm mechanisms deal with the need for rapid reactions using fast pattern recognition based on information from many sources, internal and external.

An alarm mechanism is likely to be fast and stupid, i.e. error-prone, though it may be trainable.

Cogaff is a schema not an architecture: a sort of ‘grammar’ for architectures

Different organisms, different artificial systems, may have

- **different components of the schema**
- **different components in the boxes**
- **different connections between components**

E.g. some animals, and some robots have only the reactive layer (e.g. insects, microbes).

The reactive layer can include mechanisms of varying degrees and types of sophistication, some analog, some digital, with varying amounts of concurrency.

Other layers can also differ between species.

CogAff and consciousness

Different architectures compatible with the CogAff schema will support very different kinds of mental processes which have some connection or other with our normal notion of ‘consciousness’.

E.g. all support some form of ‘sentience’, i.e. awareness of something in the environment, including the fly’s awareness of your hand swooping down to catch it.

If two perceptual pathways are affected when the fly detects motion of the hand, one relatively slow normal behavioural control pathway, and a rapid reaction pathway involving an alarm mechanism, then the fly has two sorts of awareness of the hand.

But that does not imply that it is aware of that awareness.

Characterising the layers

The differences between the layers are complex and subtle.

Some of the differences are discussed in other slide presentations here
http://www.cs.bham.ac.uk/~axs/misc/talks/

Further discussion is in the papers in the Cogaff directory
http://www.cs.bham.ac.uk/research/cogaff/

It may turn out that there are better ways of dividing up levels of functionality, or that more sub-divisions should be made – e.g. between analog and discrete reactive mechanisms, between reactive mechanisms with and without chained internal responses, between deliberative mechanisms with and without various kinds of learning, or with various kinds of formalisms, and between many sorts of specialised "alarm" mechanisms.

The COGAFF schema is still a draft, likely to evolve

Architectural change in an individual

Learning can introduce new architectural components, e.g. the ability to read music, the ability to write programs.

Development of skill (speed and fluency) through practice can introduce new connections between modules, e.g. links from higher-level perceptual layers to specialist reactive modules.

For instance, learning to read fluently, or developing sophisticated athletic skills.

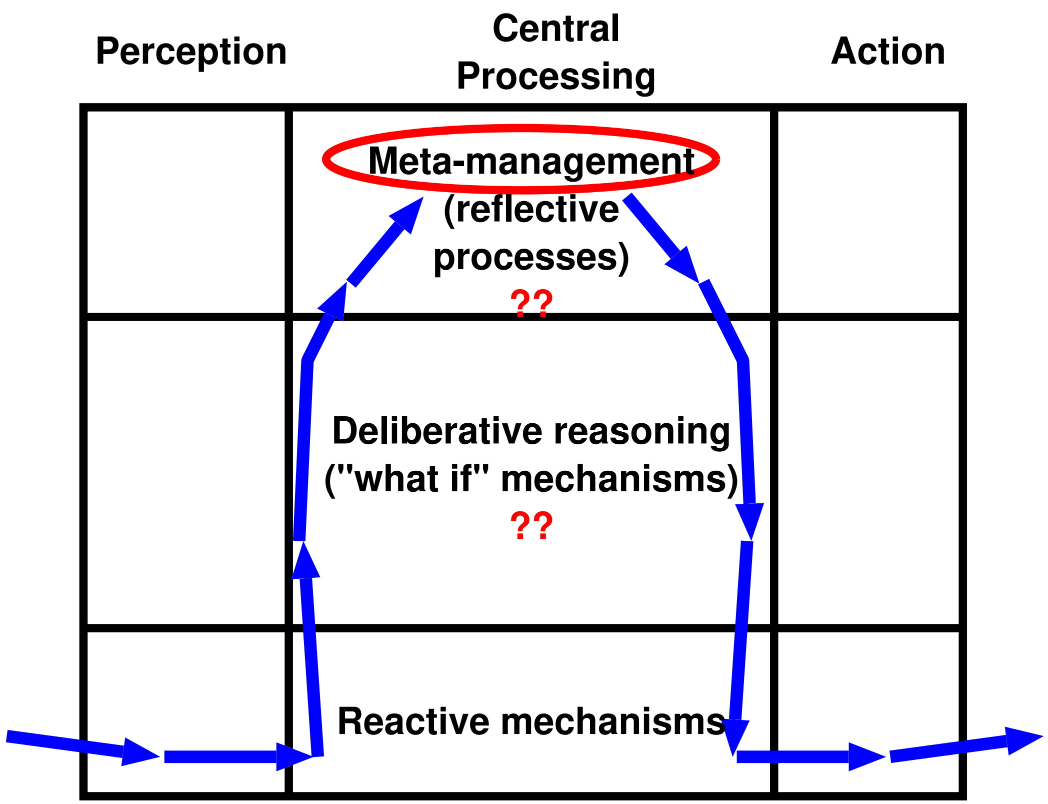
Highly trained skills can introduce new “layer-crossing” pathways, e.g. visual pathways: rapid recognition of a category originally developed for deliberation can, after training, trigger fast reactions.

So the varieties of consciousness that are possible within an individual can develop over time.

Some sub-species of the CogAff schema follow.

An example sub-category: Omega architectures

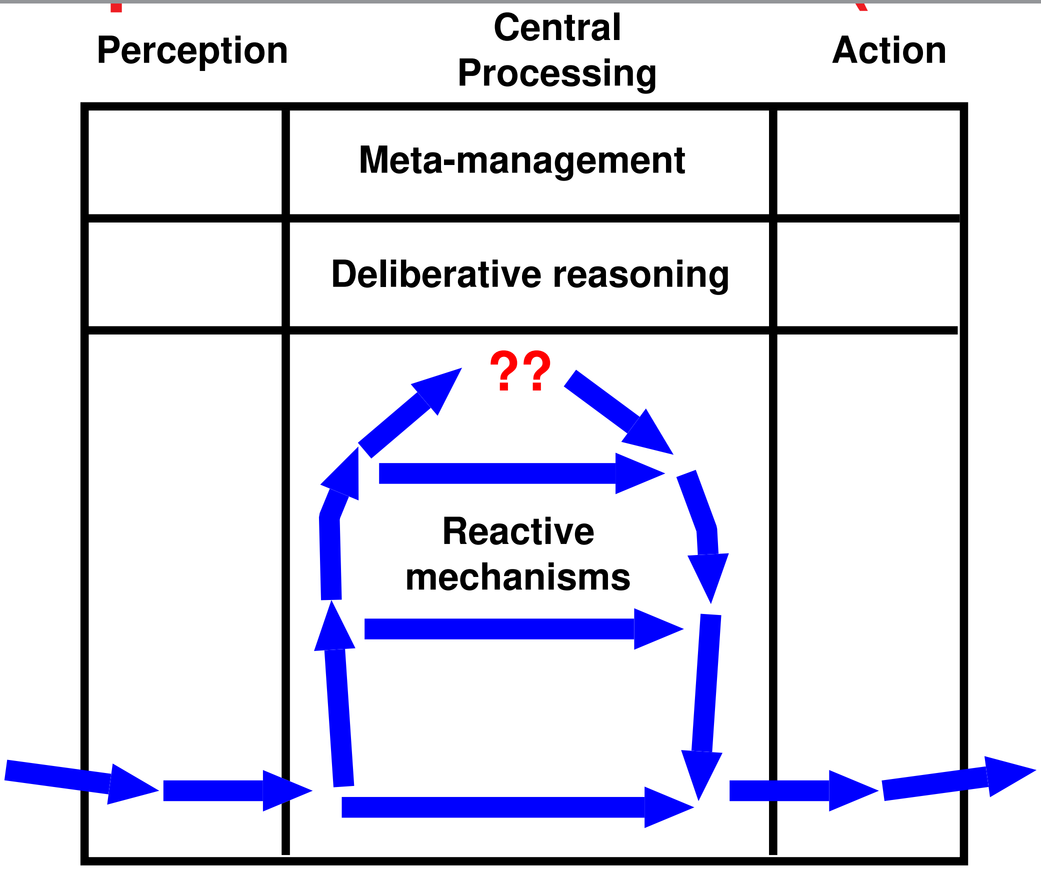
Here only some of the possible routes through the system are used, forming roughly the shape of an Omega: Ω



This is just a pipeline, with “peephole” perception and action, as opposed to “multi-window” perception and action.

E.g. Cooper and Shallice: Contention scheduling, Albus 1981.

Another sub-category: Subsumption architectures (R.Brooks)



This could be useful for certain relatively primitive sorts of organisms and robots. (E.g. Insects, fish, crabs?)

SUMMARY

We offer a solution to muddles about consciousness based on:

- Virtual machine functionalism – emphasising internal causal powers, states, interactions, implemented in physical mechanisms, but not themselves physical (like poverty).
- Comparative studies of minds of many kinds
  infants, toddlers, children humans, healthy, damaged, disturbed, many kinds of animals, many kinds of machines.
- Investigation of the (huge) space of virtual machine architectures, including both evolved and designed architectures.
- Refine and extend (not replace) existing confused concepts with several families of architecture-based concepts (many ‘elephants’)
  different information processing architectures support different varieties of consciousness – and different varieties of learning, motivation, beliefs, emotions, intentionality, etc..
- Separation of conceptual confusions from empirical issues
- Integration of multiple disciplines.
  E.g. Philosophy, psychology, ethology, neuroscience, evolution, artificial intelligence, software engineering and computer science.

Different sorts of functionalism

WARNING

- Many people write as if 'functionalism' were a simple, well defined generally understood concept.
- It is normally assumed that functional states are defined in terms of a web of possible causal connections between inputs and outputs of the whole system – i.e. crossing the physical boundary.
- Our sort of (virtual machine) functionalism (like Ryle's) refers to states of virtual machines whose causal connections can be defined in terms of the virtual machine states, e.g. like states of a Lisp or Java virtual machine.
- These states can exist and interact without any external causal connections. They could be decoupled from sensors and motors, like a program that neither reads nor prints anything, but merely explores theorems derivable from a set of axioms.
- This is nothing like behaviourism.
- It leads to puzzles about how mechanisms could evolve if they don't produce behavioural effects that influence biological fitness: answer side-effects.

A complex, long term research programme

Our approach is a mixture of science and philosophy.

The science includes the study of

Evolvable virtual information processing architectures

As part of a larger study of

– The space of possible designs,
– the space of possible niches,
– relations between those spaces,
– trajectories within those spaces,
– the dynamics of interacting trajectories in those spaces

(See talk 4 here http://www.cs.bham.ac.uk/~axs/misc/talks/)

This work can provide a general conceptual framework for defining different types of consciousness in terms of the information processing architectures that support them and the kinds of states, events and processes that can occur in those architectures.

That can then lead to empirical investigations to find out which animals have which sorts: a deeper and more rewarding quest than looking for the presence or absence of one ill-defined sort.

Is something missing?

There will always be people who are convinced that this sort of project inevitably fails to answer the questions about consciousness which *they* think are the real ones.

Often these are people who say some of the things in our earlier list of typical utterances about consciousness, e.g.

- It's indefinable, knowable only through having it.
- It's what it is like to be something (hungry, in pain, happy, a bat...).
- Zombies are possible: machines that are indistinguishable from us could lack consciousness.

The fact that many people think like this is *part of what needs to be explained* by any adequate theory of consciousness. Our explanation is that it is a side-effect of some of the processes made possible by the existence of a meta-management layer which allows an information-processing system to attend to aspects of its own internal functioning, e.g. some of the intermediate states in its sensory mechanisms.

Robots with qualia

CONJECTURE:

When robots have suitably rich internal information processing architectures some of them will also feel inclined to talk about consciousness, and qualia, in this sort of way.

NOTE: Science fiction writers thought of this long ago.

Thus even if philosophical theories about qualia, about “what it is like to be something”, involve much confusion and even error, it does not follow that they are completely wrong. They are based on a correct partial view of the nature of mind.

A meta-management system can develop its own conceptual framework for categorising its own internal states and processes, which may have features which it cannot possibly communicate (reliably) to others.

This could lead robot philosophers to raise unanswerable questions about whether others have the same experiences as they do.

Solution/Dissolution to philosophical puzzles about consciousness

When a question has no answer because it is based on muddles, the next best thing to an answer is a theory of the architectures and mechanisms which lead, in humans, to the question being formulated, and would also do in machines with a similar information-processing architecture.

But it appears to be much easier to persuade six blind men that they are feeling a small part of a much larger object.

It is important to distinguish two questions

(1) Is any information processing virtual machine architecture sufficient to produce *mental* states and processes like ours?

(2) Which, if any, of these virtual machines can be implemented on a computer (of the sort that we currently know how to build)?

It is often assumed, wrongly, that a negative answer to (2) implies a negative answer to (1).

That's because many people do not appreciate that the general notion of an information processing machine is not defined in terms of computers – computers just happen to be the best tools we currently have.
Next century we may invent new kinds of information processing engines, as different from computers as computers are from mechanical calculators.

It might turn out that certain sorts of virtual machine architectures are adequate for the implementation of all typical adult human mental phenomena, but that no digital computer is able to support them all.

Finding out the answers requires us first to clarify meanings of the questions and the available answers. I know no better method than the method outlined here.

Is something left out?

Some people feel that the kind of explanation being offered here cannot suffice since they are *convinced* that something is left out.

Often this takes the form of the ‘zombie’ argument: a robot could have all the information processing capabilities described here and still lack “this” – said attending inwardly, using human meta-management capabilities. I.e. the robot might be a ‘zombie’.

(For a survey of arguments see D.J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, 1996)

Usually this is based on a confusion between (a) a robot having all the *externally observable* behaviours humans have and (b) a robot having all the *internal* information processing capabilities humans have.

Case (b) is hard to understand if you are not familiar with virtual machines!

When internal processing of a human-like virtual machine is described *in great detail*, including the meta-management abilities involved in thinking about qualia, it is not clear that anything intelligible is left over: the zombie description becomes incoherent.

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/˜axs/ijcai01

Falsifiability? Irrelevant.

Within the proposed framework we can make simultaneous progress in science (several different sciences) and philosophy, including investigating relationships between brain mechanisms and the virtual machine architectures described here.

Is the theory falsifiable? — That's the wrong question:

What's more important than (immediate) falsifiability is the ability to generate large numbers of different, non-trivial consequences about what is possible, e.g. implications about possible types of learning, about possible forms of perception, about possible types of emotions.

You can't empirically refute statements of the form "X can happen".

But they can open up major new lines of research and unify old ones.

Following Popper and Lakatos we need to ask whether this will turn out to be a progressive or a degenerative research programme.

Why not help us to find out?

This talk presented only a subset of the concepts and theories we have been developing

There's more in online slide presentations (postscript and PDF):

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

See also

http://www.cs.bham.ac.uk/research/cogaff/
(papers)

http://www.cs.bham.ac.uk/research/poplog/freepoplog.html
(software tools for exploring hybrid architectures)

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 work of many well known philosophers, and developed with the help of many students, colleagues and friends, including Margaret Boden, Ron Chrisley, Pat Hayes, Steve Allen, John Barnden, Luc Beaudoin, Catriona Kennedy, Brian Logan, Riccardo Poli, Matthias Scheutz, Ian Wright.

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

David Booth reminded me of the importance of value judgements in these discussions.