Slide presentation

Architectures for Emotional Animals and Machines

These slides were presented at a course on Artifificial Intelligence/Cognitive Science held in Germany 4th–5th March 2002.
They are now being made available as background information for the discussion on Artificial Intelligence to be broadcast on BBC2 at 30 minutes past midnight on Thursday 14th March (00:30am on 15th March) 2002.
This is part of the BBC/Open University series, The Next Big Thing.

See http://www.open2.net/nextbigthing

Architectures for
Emotional Animals and Machines

Course of four lectures presented in Germany March 4-5th 2002
See http://www.cs.bham.ac.uk/˜axs/ik2002 for details

Aaron Sloman

http://www.cs.bham.ac.uk/˜axs
School of Computer Science
The University of Birmingham, UK

With much help from

Luc Beaudoin, Ron Chrisley, Catriona Kennedy, Brian Logan, Matthias Scheutz, Ian Wright, and other past and present members of the
Birmingham Cognition and Affect Group
and many great thinkers in other places

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

Free software tools can be found here:

http://www.cs.bham.ac.uk/research/poplog/freepoplog.html

Provisional plan

- **Section 1:** Virtual machine architectures and their relations to physical machines. Minds and their relations to brains. Muddled ideas about consciousness and emotions. Architecture-based concepts are better. What is an architecture? How many kinds are there?

- **Section 2:** Reactive, deliberative and meta-management layers in the human architecture, and how, and why, they might have evolved. (Partial critique of those distinctions.)

- **Section 3:** How the layers relate to different sorts of emotions, moods, and othr affective states and processes.
  The COGAFF architecture-schema and the H-Cogaff proposed schematic architecture.

- **Section 4:** What are YOUR objections? How does this relate to alternative theories?
  Have we accounted for everything (including qualia?).
  Unsolved problems: where next?

"Advertisement"

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/

The machines run for weeks or months without a crash

Overview (1)

This course attempts to explain a new way of thinking about philosophy, psychology, neuroscience, and AI in terms of

– **evolvable**
– **architectures**
 for
– **information-processing machines**
– **especially virtual machines.**

Consider the variety of inconsistent definitions of "emotion" researchers propose.
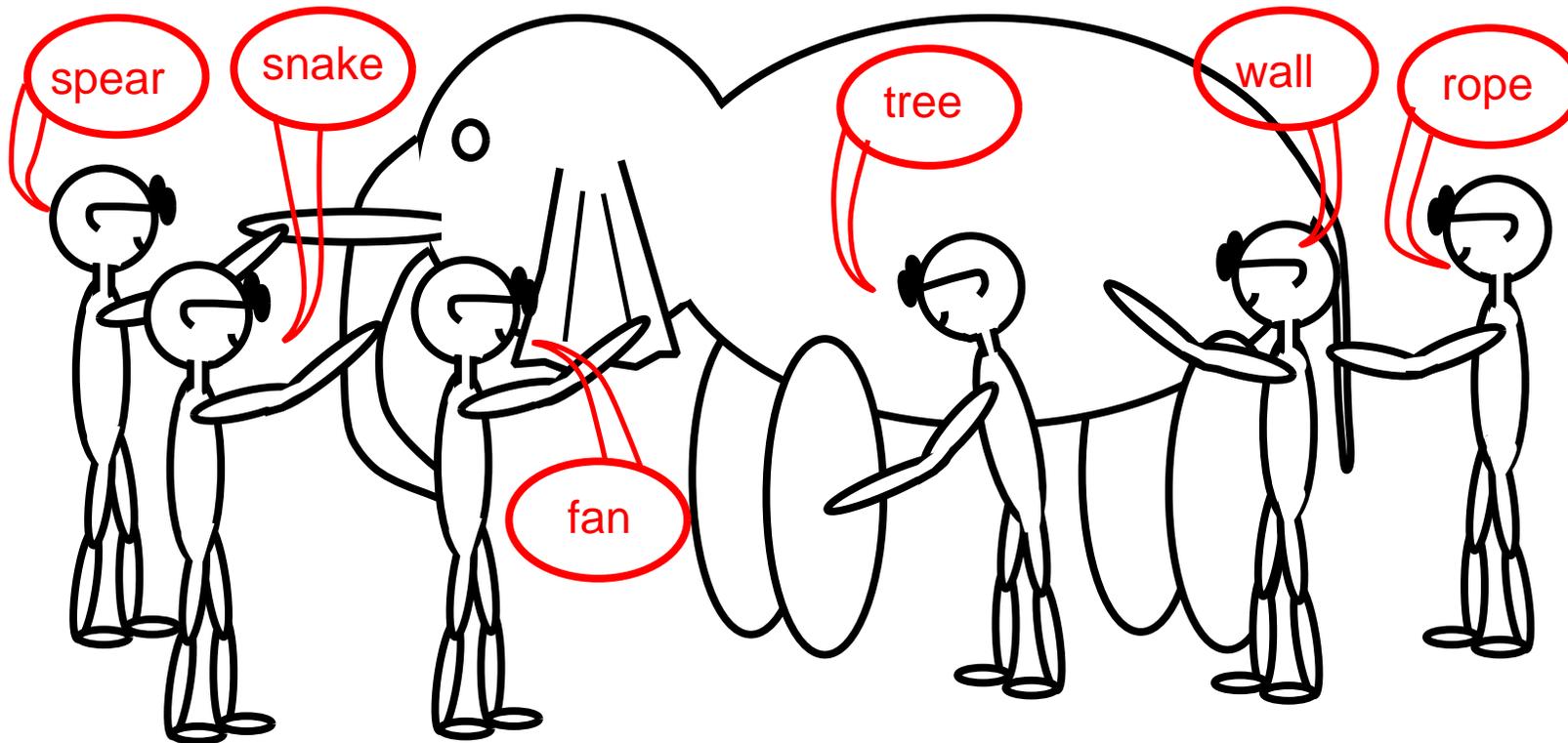
Likewise with "consciousness".

This is due to widespread conceptual muddle

Perhaps it is like six blind men describing an 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?

...continued

We can hope to see "the whole elephant" more clearly if we understand the variety of processes that can occur within a human information processing architecture.

Even more importantly:

**by looking at different architectures, for human adults, for children, for dogs, for rats, for fleas....
we may understand the even larger variety of affective states and processes that different architectures support**

There are many "elephants" for us to study.

Many other mental concepts:

"CONSCIOUSNESS", "BELIEF", "INTENTION", "INTELLIGENCE", "PLEASURE", "PAIN", "FREEDOM", ETC.

can be clarified in an architectural framework.

E.g. we can better understand our notion of "autonomy" if we see it as architecture-based: we can then understand the diverse types of autonomy (or free-will) supported by different architectures.

How do you find out what the architecture is?

Studing the architecture of a complex system is much easier if you have designed it yourself. You then know what the important parts are, how they interact, how they develop, etc. (Though sometimes we design things that are too complex for us to understand.)

Trying to understand a naturally occurring architecture, e.g. the architecture of a human mind, can be very difficult, since just observing a system from the outside will not tell you how it works.

Different information processing architectures can be produce exactly the same input/output relationships. So we have to use many kinds of evidence, including

➜ knowledge gained from neurosciene about the physical mechanisms used

➜ knowledge gained from AI about which algorithms, forms of representation and architectures are good for which purposes,

➜ introspective knowledge about what sorts of thoughts and feelings we can have

➜ knowledge about biological evolution which may constrain the types of information processing architectures to be found in living organisms.

Producing a good theory about the architecture, like all deep science, is a speculative creative process: there are no rules for theory construction. But we can compare merits of rival theories.

Benefits of the architecture-based approach

Construing familiar concepts of mind as *architecture-based* can give us new, deeper insights into what we are and how we work and, for those who so wish, a better basis for designing human-like synthetic agents – e.g. for entertainment purposes.

For instance, we'll see that
- different perceptual processes
- different types of decision making
- different types of learning
- different sorts of emotions
  (primary, secondary, tertiary, ...)

are associated with different architectural layers.

What is an architecture?

Roughly an architecture is whatever is common to two complex entities that are similar insofar as they have

- similar parts
- connected in a similar way

An architecture is a kind of abstract specification for something complex, whether it is a building, a university, a railway system, a physical computer, an operating system, a symphony or a mathematical proof.

An architecture can have instances. Instances of the same architecture will share a common structure, though they need not be exactly alike.

The architecture may be specified in great detail (e.g. the architecture of a house specified down to the individual bricks, nails, planks) or at a relatively abstract level (e.g. specifying the house in terms of the number of rooms, their sizes, interconnections, windows, doors, etc.)

An architecture can have instances. Instances of the same architecture will share a common structure, though they need not be exactly alike, e.g. two houses with the same architecture filled with quite different furniture and painted different colours.

Some important spaces

- **We study not just one architecture but the space of possible architectures – "design space".**

  We can talk about a *design* without presupposing a *designer*.

- **We relate architectures to sets of requirements – i.e. to niches.**

- **So design space is related to "niche space".**

- **During co-evolution of species, different designs and different niches constantly change and constantly interact.**

  Design of species X affects the niche of species Y, and vice versa.

  X's niche may cause X's design to change, altering Y's niche, and therefore Y's design.

  Evolution involves multiple interacting trajectories in design space and in niche space.

Relations between designs and niches

Don't think of it as a simple numerical fitness function, or a total ordering.

→ **A design may fit more or less well in several niches, with different advantages and disadvantages in each.**

  E.g. Some animals can survive both on forests and in open terrain.

→ **Similarly a niche may be filled by different designs, with different advantages and disadvantages.**

  E.g. two related species of birds may compete in the same terrain.

→ **See "Which?" (consumer magazine) reports for example niche/design relationships.**

  Often the best buy for one person is not the best buy for another, but trade-offs between requirements and options can be explained.
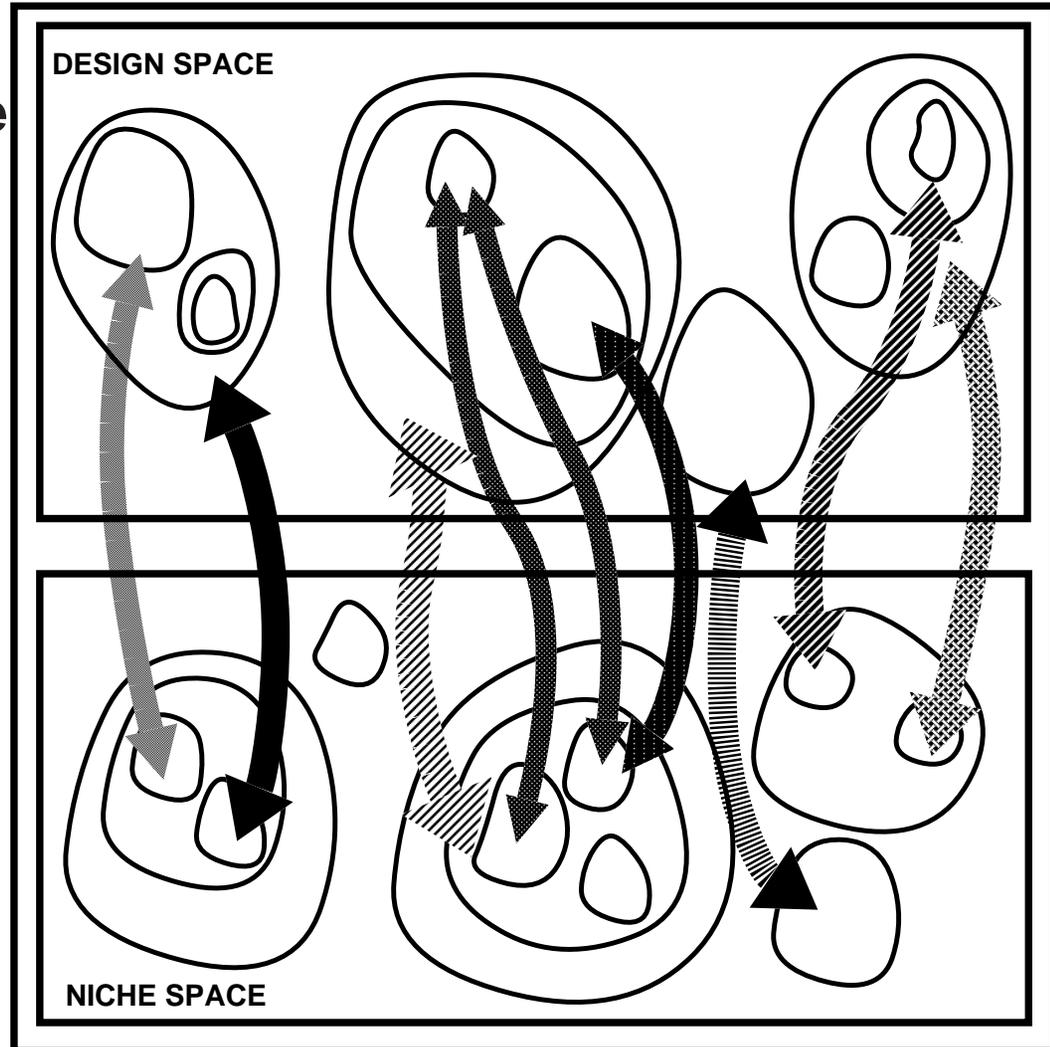
Design space and niche space

There are discontinuities in both design space and niche space: not all changes are continuous (smooth).

Many researchers look for one "big" discontinuity (e.g. between non-conscious and conscious animals).

Instead we should investigate many small discontinuities as features are added or removed.

A continuum (smooth variation) is not the only alternative to a big dichotomy.

DESIGN SPACE

NICHE SPACE

Trajectories in design space and niche space

There are different sorts of trajectories in both spaces:

– **i-trajectories:**

    Individual learning and development

– **e-trajectories:**

    Evolutinary development, across generations, of a species.

– **r-trajectories:**

    Repair trajectories: an external agent replaces, repairs or adds some new feature. The process may temporarily disable the thing being repaired or modified. It may then jump to a new part of design space and niche space.

– **s-trajectories:**

    Trajectories of social systems.

Some e-trajectories may be influenced by cognitive processes (e.g. mate-selection). We can call them c-trajectories
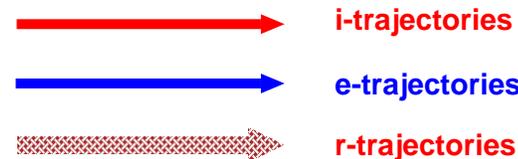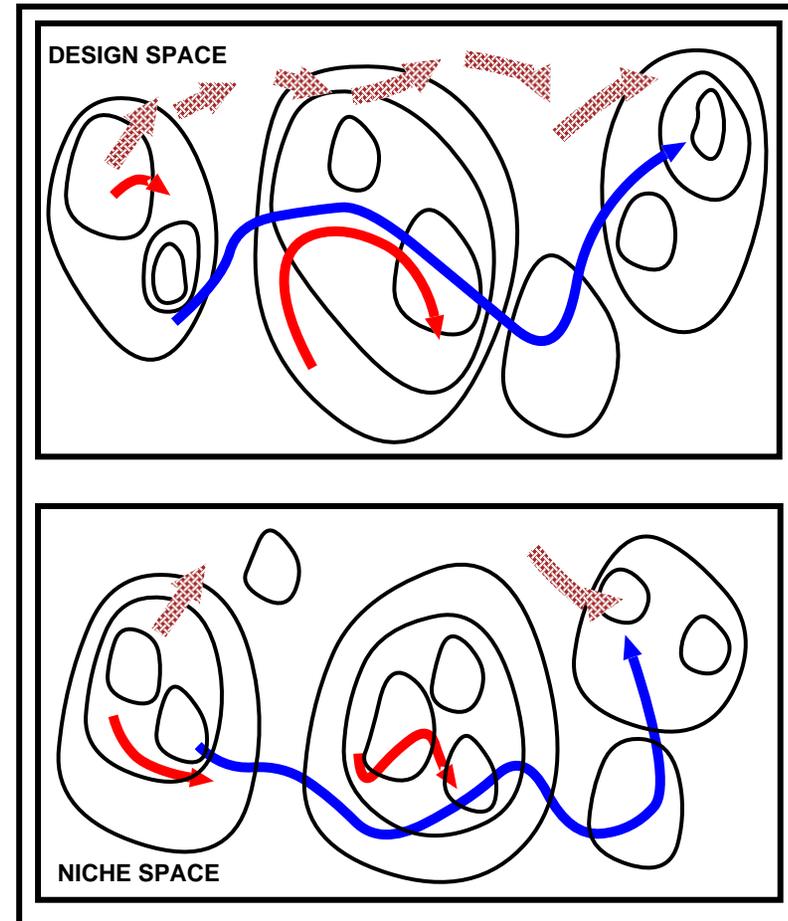
A COROLLARY:
Supposed dichotomies become complex taxonomies.

EXAMPLES:

- **We may think we understand what a 'reactive' system is, but when we investigate closely we find a space of types of reactive systems with importantly different properties.**
- **Likewise the category of 'deliberative' systems divides into sub-categories when we study different architectures.**

Demonstrations presented during the lectures:

- **Reactive systems**
  - **Flocking behaviour** Blindly following a "leader" by reacting only to sensory input
  - **Emotive reactive system** emotional states produced by different percepts alter behaviour
  - **The sheepdog demo** It has different global states with different collections of reactions
  - **Eliza** A reactive system where reactions include instantiated variables

- **A deliberative system**
  - **A blocks world conversationalist** Loosely modelled on Winograd's SHRDLU (1971)

Exploring architectures and their implications teaches us to abandon simple classifications of systems, and simple classifications of the processes that can occur in them.

The evolution of AI

→ **Prehistory**: **Golems, Frankenstein stories, mechanical toys.**

→ **Logic, Turing machines, philosophy of computation**
  (20th Century up to about 1950)

→ **Early experiments: GPS, Neural nets, Logical engines, games**
  (up to 1970)

→ **Robotics, NLP, Concept formation, Expert systems, HYPE**
  (1970s)

→ **Neural nets, HYPE**   **(1980s)**

→ **Hybrid systems**   **(1990s)**

→ **The dawn of ideas about architectures**   **(Mid 1990s)**
  Realisation that it's not enough to investigate
  – **Representations**
  – **Algorithms**
  – **The knowledge needed**
  We also need to understand how to put things together:
    I.e. WE NEED TO STUDY ARCHITECTURES.

    (Minsky, Brooks, Nilsson, others.)

Contrast "Heterarchy" in the early 1970s:

In the early 1970s there was vogue in AI for processing that was not hierarchic but heterarchic:

● Multiple types of system interacting in a non-predetermined sequence.
(E.g. SHRDLU, MIT vision demo, ...)

● But it was still all sequential, with a single, but changing locus of control.

● Compare neural nets: distributed but still unitary locus of control.

An architecture with a collection of distinct mechanisms operating concurrently and cooperatively on different tasks and subtasks can be more robust than a single process, e.g. because one can detect and compensate for failings in another.

Contrary to popular opinion this kind of robustness can be implemented on computers, using multi-processing operating systems.

What sorts of architectures?

- **Information processing architectures**
- **Made of many components, all concurrently active**
- **Using many different sorts of information-bearing structures**
  - discrete vs continuous
  - fixed vs variable structure/complexity
  - compositional vs non-compositional semantics
  - Neural, symbolic, other ....
- **Manipulating those structures in many different ways**
  - Simple homeostatic control loops
  - Hierarchic control systems
  - Neural learning and control systems
  - Chemical control systems
  - Rule-based systems (parsers, theorem provers, rule-execution mechanisms)
- **For many different purposes, on different time-scales:**
  (Individual actions, learning, development, evolution, social evolution.)

Virtual machine architectures

What are physical machines?

– Machines studied in the physical sciences (physics, chemistry, geology, meteorology, astronomy, cosmology, ...)

– Concerned with manipulation of matter and energy

What are virtual machines?

All the many kinds of machines that are implemented in, or supervene on physical machines, but are not themselves physical, e.g. because their ontology and laws of behaviour are not those of physics.

Especially information processing machines.

The interesting ones are virtual machines performing abstract operations on abstract entities.

Should we be worried about spooky non-physical entities?

In 1947 Gilbert Ryle published a book *The Concept of Mind*, in which he poked fun at the theory of a mind (or spirit or soul) as a kind of "ghost in the machine".

We now know that a computational virtual machine can be a sort of ghost in a machine.

Every intelligent ghost must include an information processing machine – for learning, remembering, wanting, hoping, expecting, deciding, thinking, etc.

A mind is just one of many types of abstract entity.

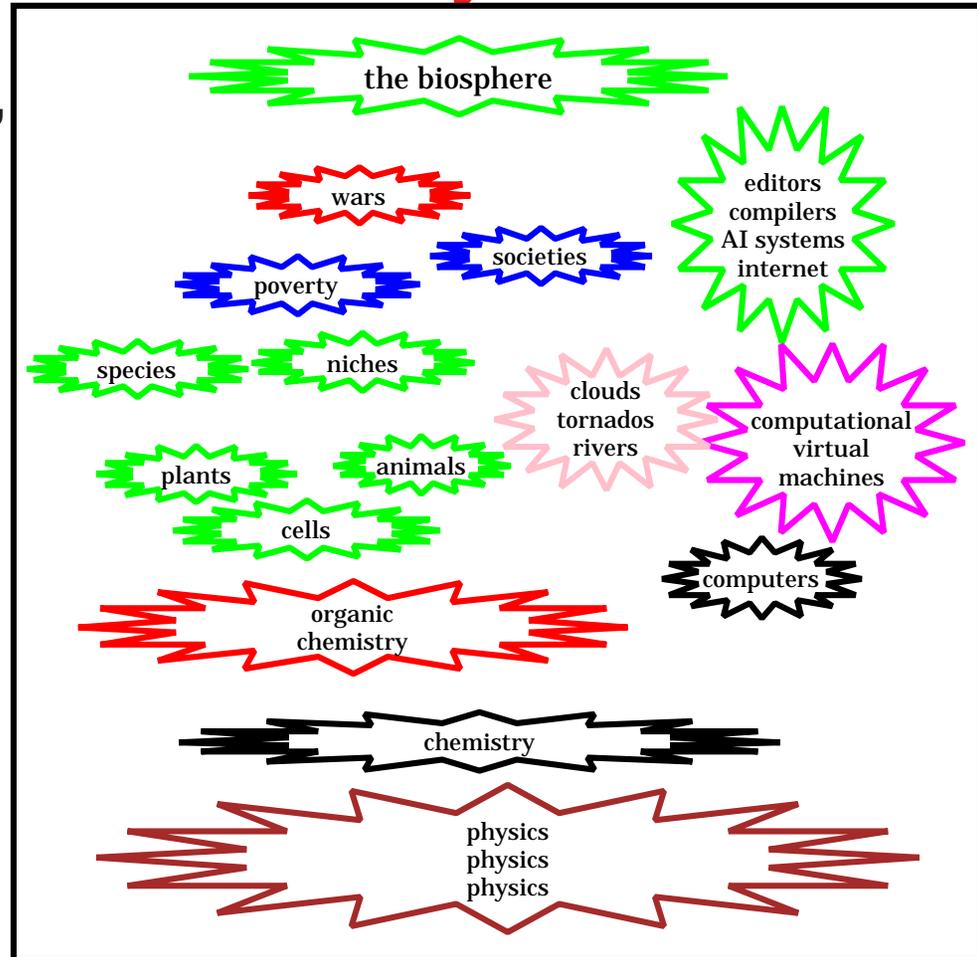Every intelligent ghost must contain a machine

Ontological levels are everywhere

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 realised (implemented) in physical systems.

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

the biosphere

wars

editors
compilers
AI systems
internet

societies

poverty

species   niches

clouds
tornados
rivers

computational
virtual
machines

plants   animals

cells

computers

organic
chemistry

chemistry

physics
physics
physics

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

Different disciplines use different approaches (not always good ones).

Methodologies for studying higher levels

- **Neuroscience:** Investigate the physiological level and try to find out how phenomena at that level correlate with behavioural and introspective phenomena.

  Mostly ignore virtual machine architectures.

- **Psychology:** Investigate relationships between environments and behaviour. Try to get data that can fit into widely used statistical packages.

  Mostly ignore virtual machine architectures.

- **Ethology:** Investigate different animals, but mainly describe them at a common-sense level. Mostly ignore virtual machine architectures.

- **Psychiatry:** Try to identify pathologies, and experimentally investigate ways of reducing or removing or preventing them. Look for causes in the environment or the brain.

  Mostly ignore virtual machine architectures.

- **Philosophy:** Defend abstract dichotomies or simple taxonomies (rational/irrational, conscious/non-conscious, action/behaviour, language/non-language, percept/belief/desire/intention, etc.)

  Mostly ignore virtual machine architectures.

Example: The ontology of biology

Biology introduces several non-physical extensions to our ontology:
E.g.

- Organisms
- Reproduction
- Growth, development and learning
- Disease, injury and death
- Species and societies
- Genes and inheritance
- Information (acquired and used by individuals or by genomes)
- Evolution, etc. ....

These are non-physical in that they have properties that are not physical properties, and are not definable in terms of physical concepts, and are not observable or measurable using physical instruments (scales, calipers, voltmeters, thermometers, etc. etc.)

We normally (apart from vitalists and some theologians) assume that, just as chemical phenomena are grounded in physics, so also:

Biological objects, events, processes are
"fully grounded (realised)" in physics and chemistry,

The grounding/realisation relation

A FIRST DRAFT DEFINITION

Phenomena of type X (e.g. biological phenomena) **are**

fully grounded in,

or

realised in,

or

implemented in

phenomena of type Y (e.g. physical phenomena)

if and only if:

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

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

(b) certain entities and processes of type Y *are sufficient for* **the phenomena of type X to exist – they constitute the implementation.**

(The actual implementation is not necessary: there can be alternative implementations.

Mutiple realizability is common in virtual machines.)

Example:

If computational virtual machines are fully grounded in (realised in, implemented in) physical machines, then

(a) computational machines cannot exist without being embodied
No word processors in a platonic heaven?

(b) their physical embodiments suffice for their existence
(Unless the software comes from microsoft?)
− no extra independent stuff is needed
− no computational spirit, soul, etc. helps the implementation (apart from virtual machines).

Computer engineering explains *how* it suffices − a non trivial problem.

But how can an abstract machine DO anything?

Causation in virtual machine ontologies

Our common sense view of the world includes 'high level' ontologies that involve *causal interactions* between components of a virtual machine. E.g.

➔ In a compiler, a parser could interact with

   an error handler and

   a code generator,

   among other things.

➔ In an operating system the process scheduler

   interacts with memory manager

   and interrupt handler, among other things.

All this talk of "interaction" presupposes a notion of "causation".

Analysing the concept of causation is probably the hardest unsolved problem in philosophy.

[[Humean theories of causation won't suffice – saying that X causes Y is not just saying that 'X precedes Y' is an instance of some true generalisation. The link with counterfactual conditionals (what would have happened if) goes beyond true generalisations.]]

Causation and Computer Science

Causation is irrelevant for (most) MATHEMATICAL computer science

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

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

➜ E.g. a Gödel number encoding a sequence of Gödel numbers can be regarded as a computation. It is a timeless, static model that accurately reflects all mathematical properties of the computation.

➜ Talking about 'time' in this context is just a matter of talking about position in a (possibly infinite, possibly branching) ordered set.

➜ State transitions are then not things that happen in time, but relations between adjacent components in the ordered structure.

➜ The notions of space complexity and time complexity in theoretical computer science refer to purely syntactic properties of a 'trace' of a program execution: another mathematical structure.

Perhaps we should say:

Computer science does not study computations, only mathematical models of computations.

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

Implementation as a mathematical relation

To say that a system X is mathematically implemented in or realised in Y is simply to say that a structural mapping exists.

Either

- X just is a subset of Y

  or

- There is a mapping from the whole of X to some part of Y which preserves certain relationships. In particular the ordering relations called state transitions in X are mapped into suitable relations in Y.

This is just a structural mapping: like the mapping of the set of positive integers onto the set of multiples of 5.

In this sort of implementation

- there is no causation
- nothing happens: only mathematical structures exist
- No energy is used.
- The issue of *reliability* does not arise, as it does for physical implementations.

Notes on mathematical implementation of X in Y

➜**X may be able to be mapped into Y in more than one way.**
(Multiple realisability)

➜**The relationship is sometimes symmetric:**
X and Y are implementable in each other (e.g. a Turing machine and a production system interpreter) if mappings exist in both directions.
For a *working* implementation such symmetry is not possible.

➜**If Y implements an *interpreter* for X, things are rather more complex. The mapping is then not just a relationship between the two structures specified independently of their state, but 'unfolds' in a sequence of 'state transitions' within Y.**

− I.e. the mathematical model needs to represent different sequences of states of X and Y and the mappings between them, instead of being simply a mapping between X and Y.

− If X is a non-terminating program which can be affected by external input, the model will have to take account of different possible inputs at different times, leading to mappings between branching infinite sequences.

− However, the code for the interpreter abbreviates all of that, in something like the way axioms and inference rules abbreviate a set of proofs generated by those axioms and rules.

Causation in Computer Applications

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

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

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

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

Software engineers want to make things happen

- **Some of what happens is in the virtual machine**
- **Some of it happens in the environment, under the control of the virtual machine**

So the engineering notion of implementation/realisation goes beyond structural mapping. It requires production of causal interactions in the virtual (implemented) machine, and usually its environment. Morevoer, energy is consumed (dissipated).

Mathematical *vs* useful implementation

In a mathematical implementation

- There is no time
- There is no causation
- No energy is consumed
- It cannot be used to change anything in the world
- It supports no counterfactual conditionals about what would happen if...
- It cannot be unreliable (e.g. subject to unknown causes of failure)

In the implementation of a useful ontology in physics

- Processes occur in time and objects endure in time
- There are causal interactions
- Energy is consumed when this happens
- There are many true counterfactual conditionals:

    e.g.

    if this pawn had been moved, that piece would have been captured,
    if this variable's value had been less than 10, the stack overflow would not have happened

- The system can (sometimes) be used to control part of the physical world
- Issues of reliability of the implementation can arise.

Concurrency and reliability

In some safety-critical systems a VM is implemented in three different physical systems which are run in parallel with constant checking that their decisions agree. If one of the systems has a flaw the chances are that the other two do not have the same flaw at the same time, so a simple vote can determine which decision is wrong.

- **From a mathematical point of view:**
  There is no difference between such a system, and a system where all three processes are run via emulations on a single time-shared CPU which is three times as fast as each of the others.

- **From the engineering point of view**
  There is a significant difference in their causal powers. The system using three physically distinct processors is more reliable, especially if the processors are separated in space. There is an even more subtle difference if the three CPUs are not fully synchronised.

We can build a mathematical model of both systems, but in order to explain why the second is more reliable we are forced to take account of additional factors, such as possibile disruptive interactions with the environment, or flaws in physical materials.

Virtual Machiness
are not defined by input-output relations

The philosophical functionalist view of mind treats mental states and entities as 'functionally' defined.

This is normally assumed to imply that they are defined in terms of some particular relationships between inputs and outputs.

BUT

– A virtual machine can run without any inputs or outputs (e.g. computing primes).

– Different virtual machines can have the same input-output relations

– The 'defining' causal relations of a VM involve *internal* states, events and processes

– The actual implementation need not have input and output transducers with sufficient bandwidth and adequate connections to check out all the distinct VM states and processes that can or do occur.

VM functionalism

So a virtual-machine functionalist analysis of mental concepts is very different from the conventional functionalist analysis.

NOTE:

States that cannot be identified through input-output relationships might be capable of being identified through direct measurement and observation of the internal physical states and processes.

But in some cases this may be physically impossible without disrupting the system.

In any case, "decompiling" may be impossible in practice because of the huge search space for possible high level explanations of observed low level physical details.

Non-redundant multi-level causation

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

NOTE:

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

(e.g. death from cancer and heart disease, or being killed by several members of a firing squad).

VM causation is non-redundant: there's no way to simply remove the VM cause of effect E without changing the world in such a way that the physical world ceases to cause E.

I.e. the VM is not like an extra soldier in the firing squad, who could be removed without making any difference.

Of course the VM cause can be replaced by another at the same level: if the bishop had not captured the pawn the rook would have.

VM counterfactuals depend on a 'normality' condition.

A good implementation tends to preserve normality, e.g. by error detection and error compensation.

But there are always limits.

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

But that's not true just of virtual machines.

All mechanisms with parts interacting causally are potentially subject to malfunctions caused by disasters – whatever the advertisements may say.

Virtual Machine events as causes

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

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

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

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

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

Must non-physical events be epiphenomenal?

Consider a sequence of virtual machine events or states M1, M2, etc. implemented in a physical system with events or states P1, P2, ….

Mental events   M1 → ? → M2 → ? → M3

?   ?

Physical events   P1 → P2 → P3
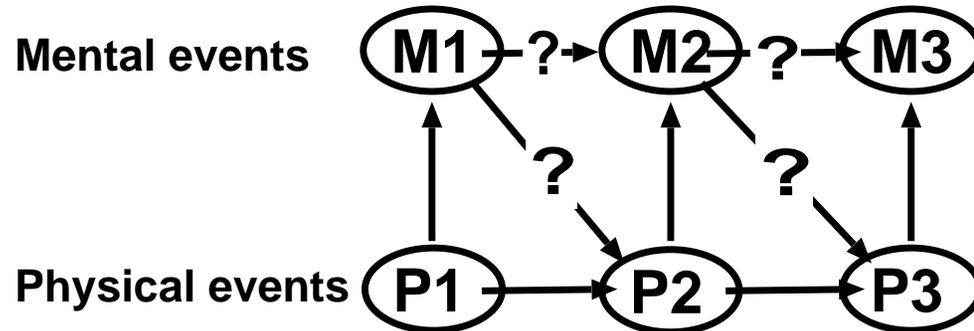
If P2 is caused by its physical precursor, P1, that seems to imply that P2 cannot be caused by M1, and likewise M2 cannot cause P3.

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

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

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

The flaw in the reasoning?

THIS IS HOW THE ARGUMENT GOES:

- **IF physical events are physically determined**

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

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

    (**I.e.** "All caused events have physical causes" IMPLIES "Nothing else can cause them: any other causes will be *redundant.*")

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

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

ONE OF THE CONJUNCTS IS INCORRECT. WHICH?

It's the second conjunct

Some people think the flaw is in the first conjunct:

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

The real flaw is in the second conjunct:

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

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

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

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

http://www.cs.bham.ac.uk/˜axs/ijcai01

Just remember: the fallacy is to treat multiple causes at different levels (e.g. VM event causing a physical event which is also caused physically) by analogy with redundant causes at the same level (e.g. two soldiers in a firing-squad killing the same victim). They are not analogous since the higher and lower level causes cannot be independently added and removed.

'Emergent' non-physical causes are possible

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

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

NOTE: Saying that the non-physical phenomena are identical with physical ones (a) does not explain anything, (b) contradicts the asymmetry in the realisation relation.

But most importantly

- **The argument that virtual machine events cannot have causal powers is based on ignorance of how actual implementations of virtual machines work, and the ways in which they produce the causal powers, on which so much of our life and work increasingly depend. (Such ignorance is far too common in our culture.)**

    More and more control systems depend on virtual machines that process information and take decisions.

- **The ideas presented here are only a first draft attempt to clarify these topics.**

- **There is much more that is intuitively understood by engineers which has not yet been clearly articulated and analysed.**

Philosophers, psychologists and neuroscientists need to learn the craft, and the underlying science, in order to avoid confusions and false assumptions.
(They also need to learn about one another's work.)

Provisional conclusion

In studying varieties of possible architectures
we must be prepared to investigate many kinds of
VIRTUAL MACHINE ARCHITECTURES

And we must expect many surprises: we still know very little about possible architectures.

Evolution has gone way beyond our current understanding.

We now attempt to gain some clarification about the varieties of possible emotions by understanding the varieties of architectures, or sub-architectures, that are capable of producing them.
In humans there are different sub-architectures, some, but not all, shared with many other animals.

Do we understand our own concepts?

Much research in AI makes use of familiar concepts, e.g. "belief", "learn", "emotion", "consciousness", "reason", "communicate", which we think we understand.

1. **A problem:** how do our concepts of mind work? We don't know, but we think we know — our ideas are riddled with confusion.

2. **A diagnosis:** part of the problem is that we are like the six blind men trying to describe an elephant.

We unwittingly use "cluster concepts": concepts that refer to a wide collection of features and relationships relevant to the applicability of the concept, where no fixed subset (or boolean combination) precisely defines the concept.

People focus on different subsets, thinking they have grasped the whole thing.

3. **A pointer:** to a good way forward: refine and extend our concepts by basing them on theories of the architecture supporting the phenomena. (Compare physical concepts and the architecture of matter.)

> ## But what sort of architecture?

Definitions of "emotion"

The psychological and philosophical literature contains many very different definitions of "emotion", e.g. in terms of

➜ Environmental eliciting conditions

➜ Typical observable behaviours

➜ Physiological measures of emotional state

➜ Brain mechanisms thought to produce them

➜ Introspective features (what it is like to feel sad, happy, etc.)

➜ Their social significance

➜ etc.

The definitions lead to inclusion and exclusion of different states and processes under "emotions".

Yet people have the impression they are all asking the same question when they ask whether machines can have emotions.
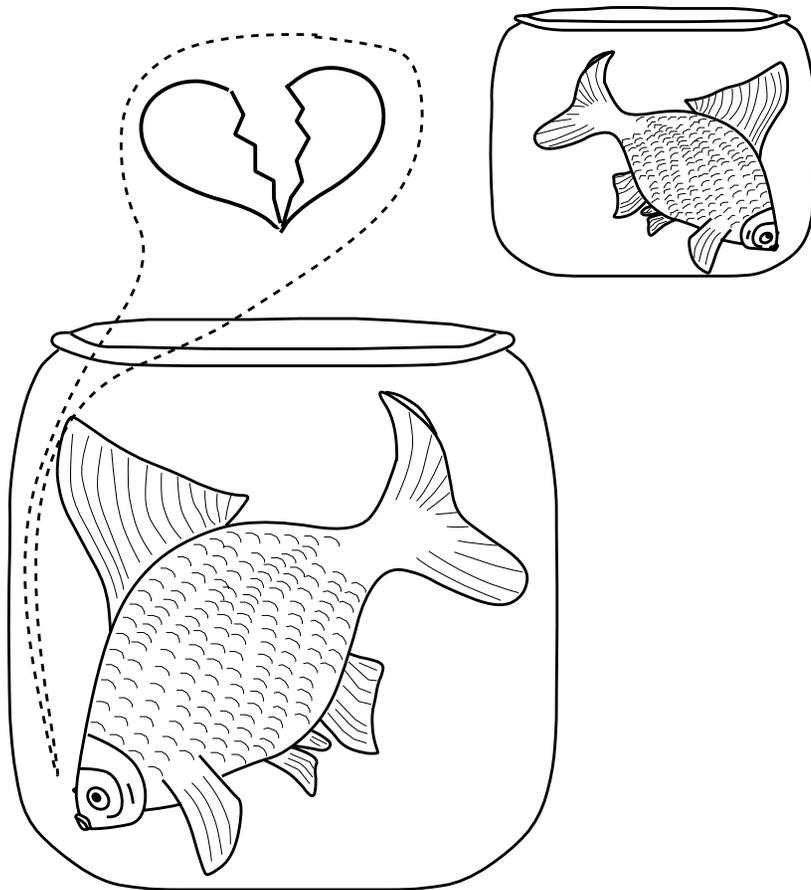
Can we find the right definition?

> **There is no right definition:**
>
> **our ordinary concepts are too muddled.**

Instead we explore classes of architecture-based concepts that more or less closely correspond to various familiar uses of words that refer to emotions.

Why can't a goldfish long for its mother?

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.**

Could information processing mechanisms suffice?

Do they suffice to produce thoughts, feelings, emotions, learning, etc.?

Yes in principle
if they have the right causal powers.

Ignore intuitions that computational processes are inadequate for the task.

The intuitions are ill-founded!

Types of architectures: a partial survey
Could the architecture be an unintelligible mess?

Some people argue that we cannot hope to understand products of millions of years of evolution. They work, but do not necessarily have a modular structure or functional decomposition that we can hope to understand.

YES, IN PRINCIPLE.

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

Problem 1: time required and variety of contexts required for a suitably general design to evolve.

Problem 2: storage space required to encode all possibly relevant behaviours if there's no "run-time synthesis" module.

Conjecture: evolution, like good engineers, 'discovered' the virtue of re-usable modules and and nearly decomposable complexes (H.A.Simon 1967).
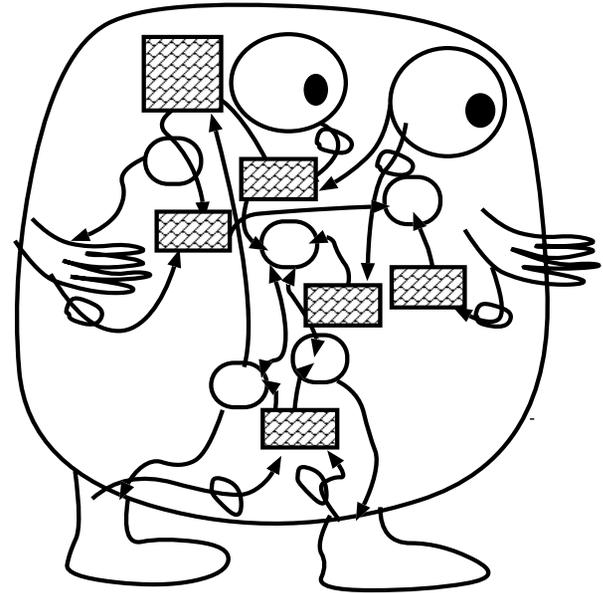
Towards a unifying theory of architectures

- **We need good general-purpose concepts for describing and comparing different classes of architectures for organisms and robots, and possibly other things.**

- **We build up our concepts by relating them to a space of possible architectures for integrated (non-distributed) agents.**

- **This space is characterised by a generic schema (a sort of grammar) specifying types of components and ways in which they may be related.**

  The schema (called CogAff) is only a tentative first draft and will certainly have to be enriched.

  (It does not cover multi-agent architectures except insofar as the components of a single integrated architecture can be viewed as agents.)

Perspectives on complete agents

1. THE "TRIPLE TOWER" PERSPECTIVE

**(Many variants – Nilsson, Albus, ...)**

"Nearly decomposable" systems. (H.A.Simon)

Boundaries can change with learning and development.

The main basis for distinguishing central from perceptual and action mechanisms: causal influence.

Perception Central Processing Action

- **The contents of the perceptual tower are largely under control of input from sensory transducers. Their function is primarily to analyse and interpret incoming information. They may also be 'in registration' with collections of sensory transducers.**

- **Similar criteria can be used for specifying contents of action tower.**

- **Contents of 'central' tower (a) change on different time-scales from those of perceptual and motor towers (b) are not closely coordinated with them.**
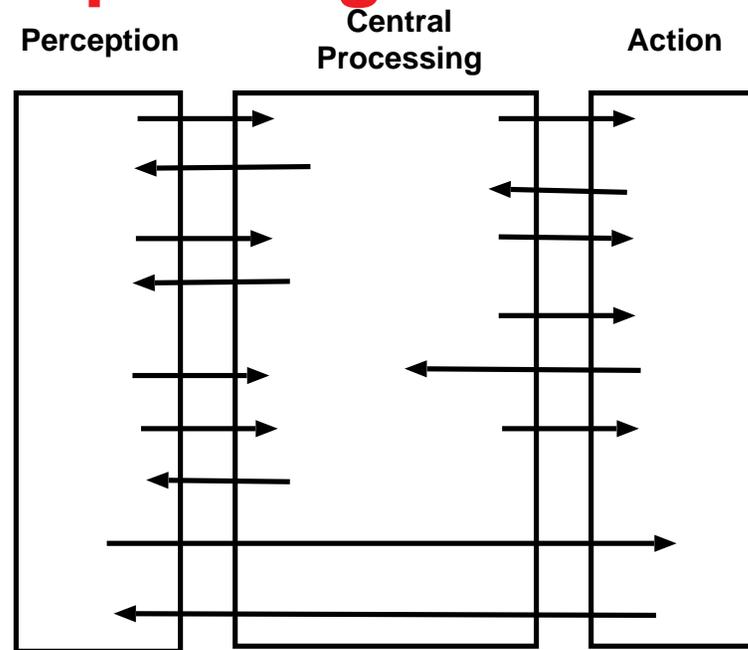
A less obvious perspective

2. THE "TRIPLE LAYER" PERSPECTIVE

Another common architectural partition (functional, evolutionary).

There are many variants – for each layer.

All mechanisms must be implemented at some level in reactive systems.

Some people separate reflexes from more complex reactive mechanisms which include internal state changes.

Meta-management
(reflective processes)
(newest)

Deliberative reasoning
("what if" mechanisms)
(older)

Reactive mechanisms
(oldest)
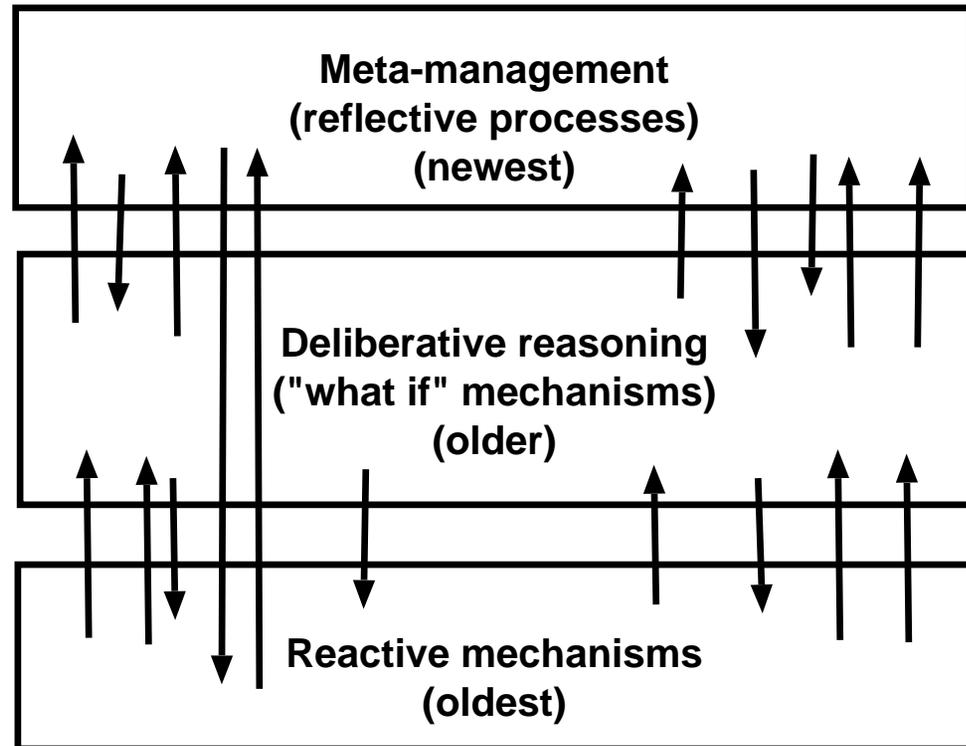
The three layers

The layers differ in:

– **Evolutionary age** (reactive oldest).

– **Level of abstraction of processing** (reactive least abstract),

– **The types of control functions, and mechanisms used**
(e.g. ability to search, evaluate, compare; amount of parallelism; use of neural vs symbolic mechanisms)

– **The forms of representation used**
(e.g. flat vs hierarchical compositional syntax and semantics)

(Many variants – for each layer)

Reactive mechanisms

➜ **They are very diverse and may include many concurrent sub-systems, including "alarm" mechanisms.**

➜ **They may include forms of learning or adaptation.**

➜ **Some reflexes (innate or learnt) connect sensors directly to motors.**

➜ **Reactive mechanisms can be highly parallel and very fast.**

➜ **They may use analog or digital components or both.**

➜ **They may include neural nets, condition-action rule systems, lookup tables, decision nets, and other mechanisms.**

➜ **Some reactions change only internal state, affecting future reactions.**

➜ **Some internal states may act as goals.**

➜ **It may be difficult or impossible to program them directly or provide explicit information for them to use (compare neural weights.)**

➜ **They make possible 'alarm-driven' primary emotions.**

NOTE: In principle any form of externally observable behaviour over any time scale can be produced by a reactive system.

However, satisfying the full set of true counterfactual conditionals required to match a deliberative system may require an impossibly large store, an imposssibly long training period, etc.

Deliberative mechanisms

➜ **Can represent and reason about non-existent or future possible entities.**

➜ **Some can also reason about what might have been the case in the past.**

➜ **They allow alternative options to be constructed, evaluated, and compared.**

➜ **They can vary in the representational forms they use and the sophistication of their semantics.**

  – **Simple deliberative mechanisms may use only one step lookahead, and very simple selection mechanisms.**

  – **More sophisticated versions use compositional semantics in an internal language.**

➜ **They require a re-usable general purpose working memory (garbage collectable)**

➜ **They require stored generalisations about what actions are possible in particular situations, and about the consequences of actions.**

➜ **They may be able to learn**

➜ **They benefit from perceptual systems that produce high-level chunked descriptions of the environment**

➜ **They may be able to train reactive systems that cannot be directly modified.**

➜ **Typically slow, serial, resource limited. (Why?) May need attention filter.**

➜ **They make possible secondary emotions using global 'alarm' mechanisms.**

Meta-management mechanisms

→ **They can monitor, categorise, evaluate, and (to some extent) control other internal processes – e.g. some deliberative processes, or some perceptual processes.**

→ **This includes control of attention, control of thought processes. (Control lost in tertiary emotions.)**

→ **They can vary in sophistication.**

→ **They require concepts and formalisms suited to self-description, self-evaluation**

→ **They support a form of internal perception which, like all perception, may be incomplete or inaccurate, though generally adequate for their functional role.**

→ **The concepts and formalisms may be usable in characterising the mental states of others also.**

→ **Different meta-management control regimes may be learnt for different contexts (different socially determined "personae").**

→ **Evolution of sensory qualia: occurs when it is useful for meta-management to look inside intermediate levels of perceptual processing (why?).**

→ **If meta-management mechanisms are damaged, blind-sight phenomena may occur.**

Varieties of meta-management

→ **There may be different types of meta-management using more or less sophisticated forms of representation and processing.**

→ **They can also vary in the types of evaluation they can apply**

→ **In humans much self-categorisation and self-evaluation is socially/culturally determined.**
(E.g. feelings of guilt or sin)

→ **The existence of meta-management may provide a "niche" encouraging evolution of higher level perceptual mechanisms categorising mental states of other agents.**

→ **This may have required parallel evolution of involuntary "expressive" behaviours (Sloman 1992 on the dangers of complete voluntary control of sincerity.)**

→ **The absence of meta-management was a major factor in the fragility and incompetence of many old AI systems (e.g. they could not tell when they were reasoning in circles.)**

→ **Mechanisms for triggering and modulating meta-management processes may produce a far wider variety of affective states than scientists have so far categorised.**
(Compare novellists!)

LAYERS + PILLARS = GRID

We can combine the two views.

A grid of co-evolving sub-organisms, each contributing to the niches of the others.

This is an architectural "schema" specifying possible components and relationships between components, not an architecture.

The CogAff schema defines a variety of components and linkages

Perception · **Central Processing** · **Action**

Meta-management
(reflective processes)
(newest)

Deliberative reasoning
("what if" mechanisms)
(older)

Reactive mechanisms
(oldest)

Not all the components, and not all the communication links, need be present in all species of natural or artificial architecture.

It does NOT specify control flow, or dominance of control: many options are left open. Contrast H-Cogaff (Below)
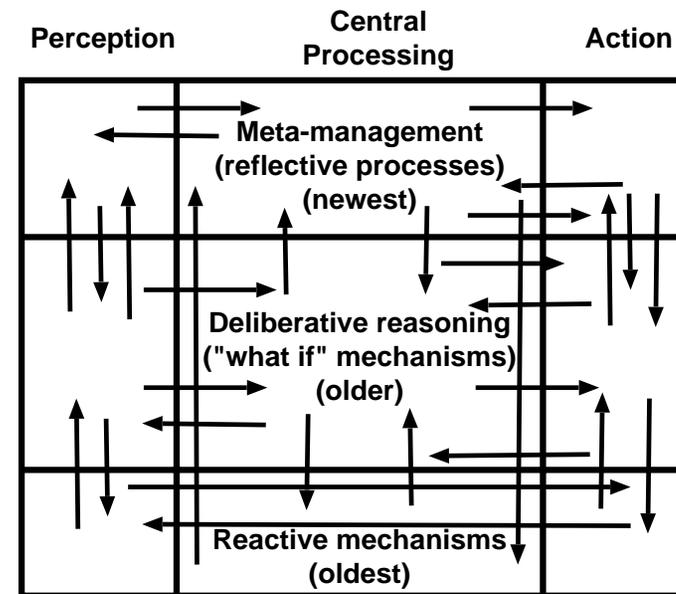
Layered architectures have many variants

With different subdivisions and interpretations of subdivisions, and different patterns of control and information flow.
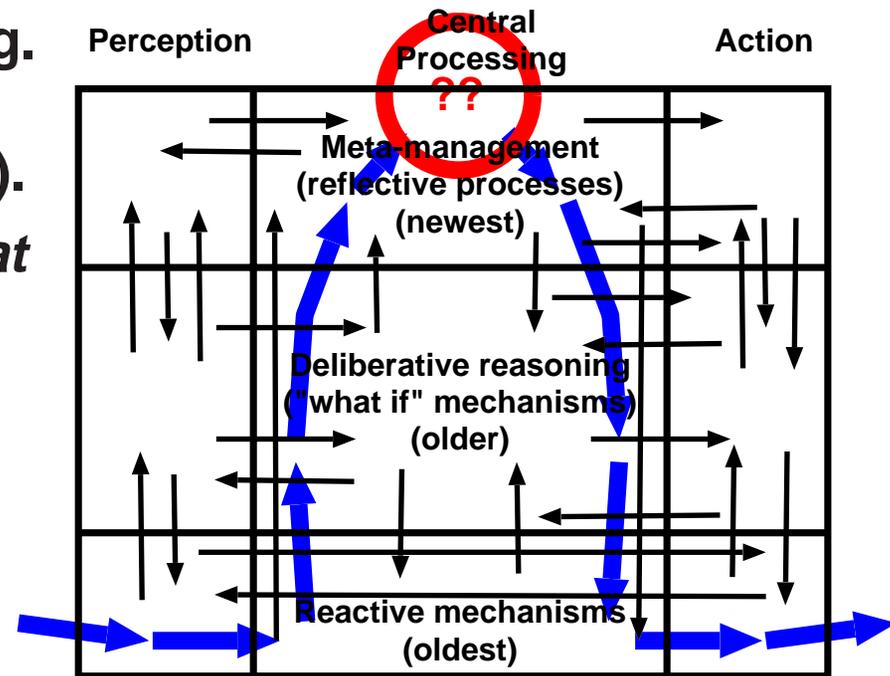
Divisions between layers can be based on:

- evolutionary stages
- levels of abstraction,
- control-hierarchy, (Top-down vs multi-directional control).
- information flow
  (e.g. the popular 'Omega' Ω model of information flow)

The "Omega" model of information flow

CogAff allows many variants, e.g. the "contention scheduling" model (Cooper and Shallice 200).

Some authors propose a "will" at the top of the omega (E.g. Albus 1981)

Rejects layered concurrent perceptual and action towers separate from central tower.

Perception   Central Processing ??   Action

Meta-management (reflective processes) (newest)

Deliberative reasoning ("what if" mechanisms) (older)

Reactive mechanisms (oldest)

What is the difference between processes in the perceptual column and processes in the central column?

Multi-level (multi-window) perception uses dedicated concurrent parsing and interpretation of sensory arrays, e.g. building new data-structures in registration with sensory arrays.

Contrast "peephole" perception.

Likewise multi-window vs peephole action.

Another special case of CogAff (Brooks): Subsumption architectures

These allow different architectural layers, but only within the reactive sub-space, where they form a sort of dominance hierarchy (unlike the layers in H-Cogaff described later.) Brooksians deny that animals (even humans) use deliberative mechanisms.
(How do they get to overseas conferences?)
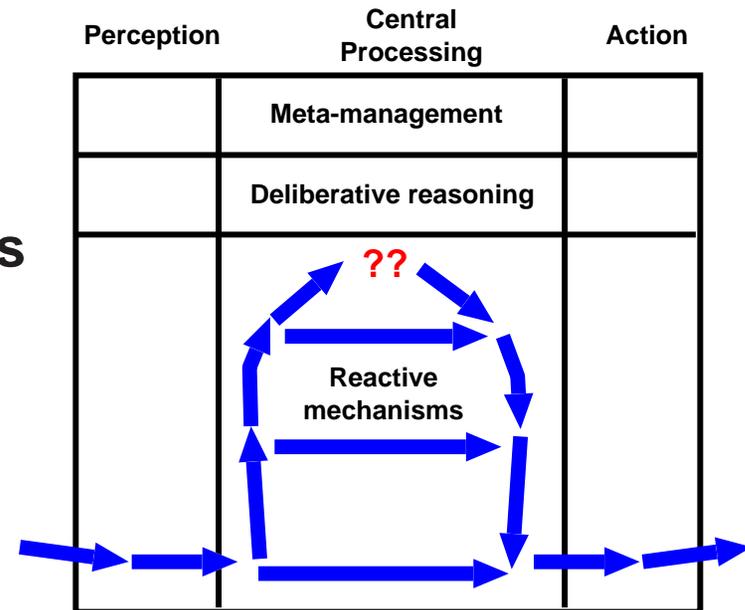
These reactive subsumption architectures are able to meet requirements for human-like capabilities ONLY IF quite unrealistic assumptions are made about evolutionary developments, storage capabilities, etc.

Perception	Central Processing	Action
	Meta-management	
	Deliberative reasoning	
	??	
	Reactive mechanisms	

Subsumption and CogAff

Subsumption, like many other architectures, uses only a subset of the mechanisms allowed in the CogAff schema.

We should avoid all dogmatism and ideology, and investigate which subsets are useful for which organisms or machines, and how they might have evolved.

The need for "alarm" mechanisms

As processing grows more sophisticated, so it can be come slower, to the point of danger. A possible remedy is to use one or more fast, powerful, "global alarm systems" (processing modulators).

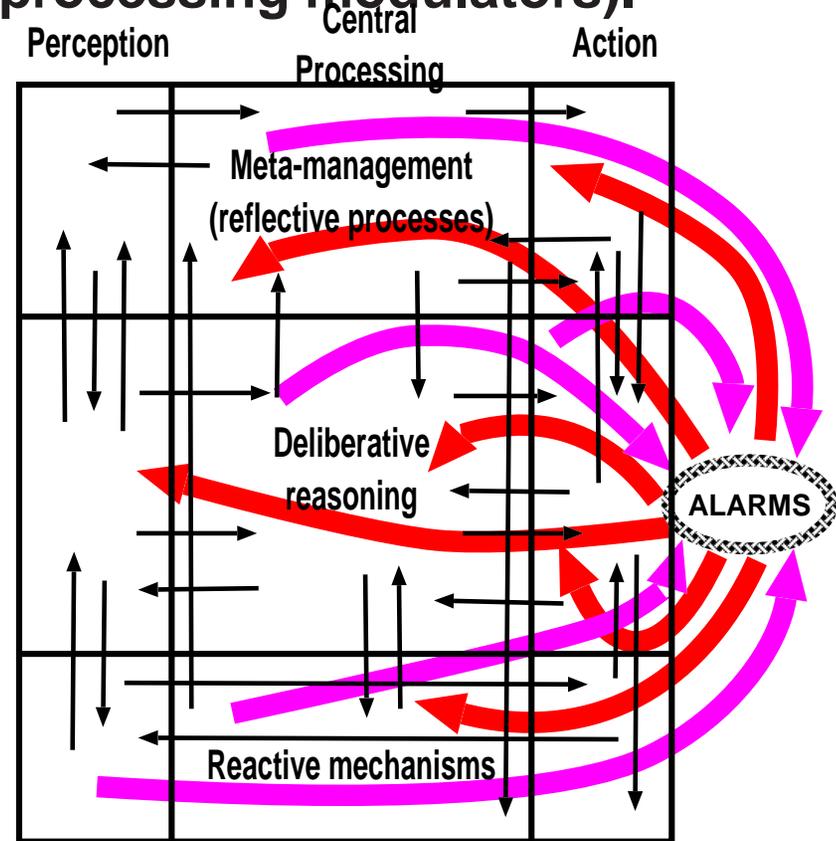ALARM MECHANISMS MUST USE FAST PATTERN-RECOGNITION AND WILL THEREFORE INEVITABLY BE STUPID, AND CAPABLE OF ERROR!

Note: An alarm mechanism is just part of the reactive sub-system. Drawing it separately merely serves the pedagogic function of indicating the role.

Many variants possible. E.g. purely innate, or trainable.

E.g. one alarm system or several? (Brain stem, limbic system, ...???) Various kinds of more or less global, more or less rapid, re-direction or re-organisation of processing.

Perception   Central Processing   Action

Meta-management (reflective processes)

Deliberative reasoning

ALARMS

Reactive mechanisms

The five Fs: Feeding, fighting, fleeing, freezing, and reproduction

Many sorts of alarms

➡ **Alarms allow rapid redirection of the whole system or specific parts of the system required for a particular task (e.g. blinking to protect eyes.)**

➡ **The alarms can include specialised learnt responses: switching modes of thinking after noticing a potential problem.**

➡ **E.g. doing mathematics, you suddenly notice a new opportunity and switch direction. Maybe this uses an evolved version of a very old alarm mechanism.**

➡ **The need for (POSSIBLY RAPID) pattern-directed re-direction by meta-management is often confused with the need for emotions e.g. by Damasio, et. al.**

➡ **Towards a science of affect:**

  Not just alarms – many sorts of control mechanisms, evaluators, modulators, mood controllers, personality selectors, etc.

Additional components are needed

A partial list is on the right:

Many profound implications regarding varieties of possible architectures, possible types of learning and development, possible effects of brain damage, varieties of affective control states.

Perception	Central Processing	Action	EXTRA MECHANISMS
	Meta-management (reflective processes) (newest)		personae
			standards
			attitudes
			formalisms
			categories
	Deliberative reasoning ("what if" mechanisms) (older)		LTM
			motives
			moods
			filters
	Reactive mechanisms (oldest)		skill-compiler

Example:

   Different sorts of learning can occur within individual sub-systems and also different sorts links between sub-systems can be learnt. (Not only links shown so far.)
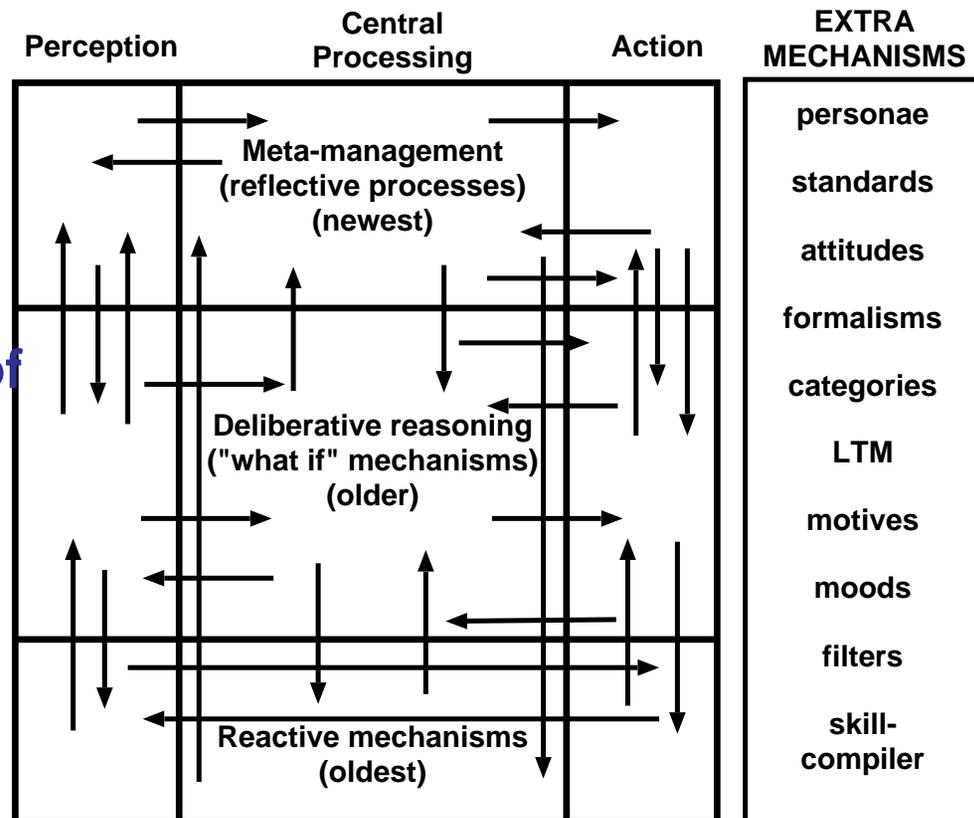
Varieties of motivational sub-mechanisms

What is motivation?

- **A type of affective control state or process, with many sub-types.**

- **Different types of contents, including bringing about, preserving, increasing, reducing, preventing, removing... some state of affairs.**

- **Motives or goals can be short term, long term, permanent.**

- **They can be triggered by physiology, by percepts, by deliberative processes, by meta-management.**

- **They can be implicit in the operation of active mechanisms, or explicit.**

- **They can operate in a totally innate (genetically determined fashion) or be learnt, or influenced by a culture (e.g. whether you enjoy eating grubs).**

- **They can be part of the reactive system, part of the deliberative system, part of meta-management.**

- **They can be implicit or explicit.**

- **They can use a wide range of representational formalisms (e.g. with or without compositional semantics).**

Motive dynamics

Generation and processing of motives requires a variety of mechanisms.

➡ There are many sorts of motive generators: MG

➡ However, motives may be in conflict,
   so motive comparators are needed: MC.

➡ But over time new instances of both may be required,
   as individuals learn, and become more sophisticated:

   ● Motive generator generators: MGG
   ● Motive comparator generators: MCG
   ● Motive generator comparators: MGC
   ● And maybe more:
   MGGG, MGGC, MCGG, MCGC, MGCG, MGCC, etc ?

There are also evaluators:

➜ **Current state can be evaluated as good, or bad, to be preserved or terminated (or intensified or reduced).**

➜ **Evaluations may interact with learning in some architectures (e.g. positive and negative reinforcement).**

➜ **These evaluations can occur at different levels in the system, and in different subsystems.**

➜ **This can account for many different kinds of pleasures and pains.**

➜ **"Error signals" form a special case**

➜ **Evaluations are often confused with emotions.**

Levels in perceptual mechanisms

Seeing the switching Necker cube requires geometrical percepts.
Seeing the flipping duck-rabbit uses far more subtle and abstract percepts, going beyond geometric and physical properties. (Compare Marr on vision)

**Necker cube**   **Duck-rabbit**
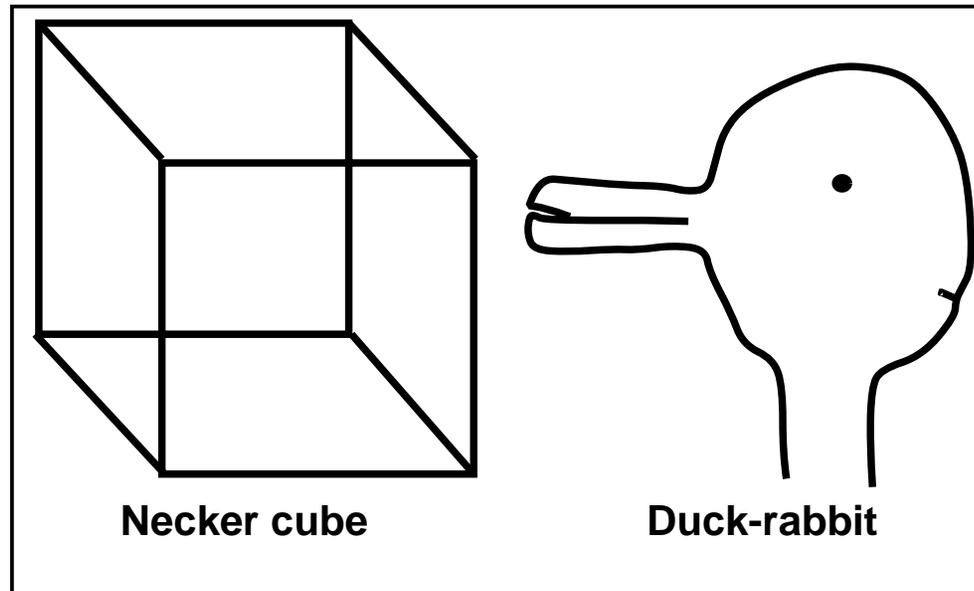
Things we can see besides geometrical properties:
- Which parts are ears, eyes, mouth, bill, etc.
- Which way something is facing
- Whether someone is happy, sad, angry, etc.
- Whether a painting is in the style of Picasso...

seeing faces

Seeing facial expression as we do may just be a very old and simple process in which features of the face trigger reactions in a pattern-recognition device.

Or it may also involve deployment of sophisticated concepts that developed only through the evolution of meta-management.
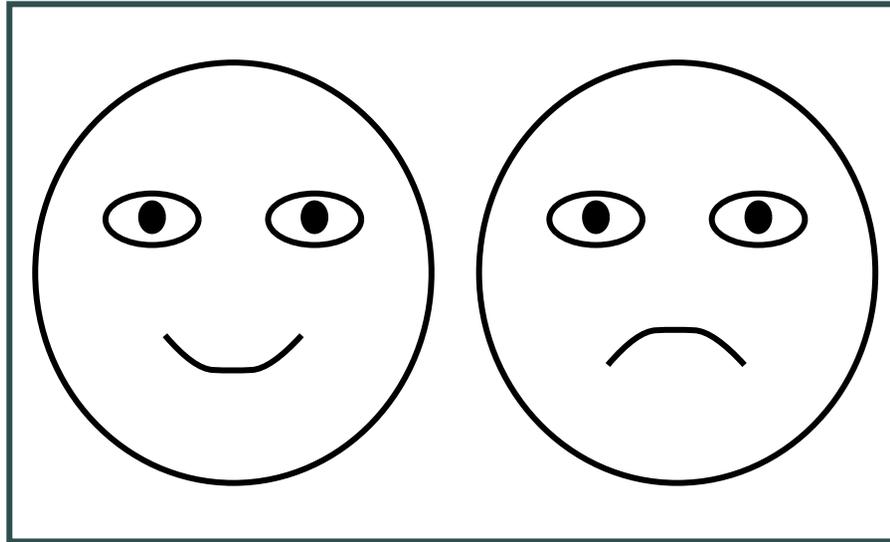

Some people see one pair of eyes as "looking happy" while the other pair "looks sad" or "looks angry".

For more on levels in perceptual mechanisms see talks on vision and visual reasoning here: http://www.cs.bham.ac.uk/˜axs/misc/talks/

Not all parts of the grid are present in all animals

Not all organisms, and certainly not all useful robots will have all the components allowed by the CogAff schema. Consider how to design an insect including an alarm mechanism?

Even reactive systems may require perceptual mechanisms to operate at different levels of abstraction, e.g. recognising food, mates, danger.
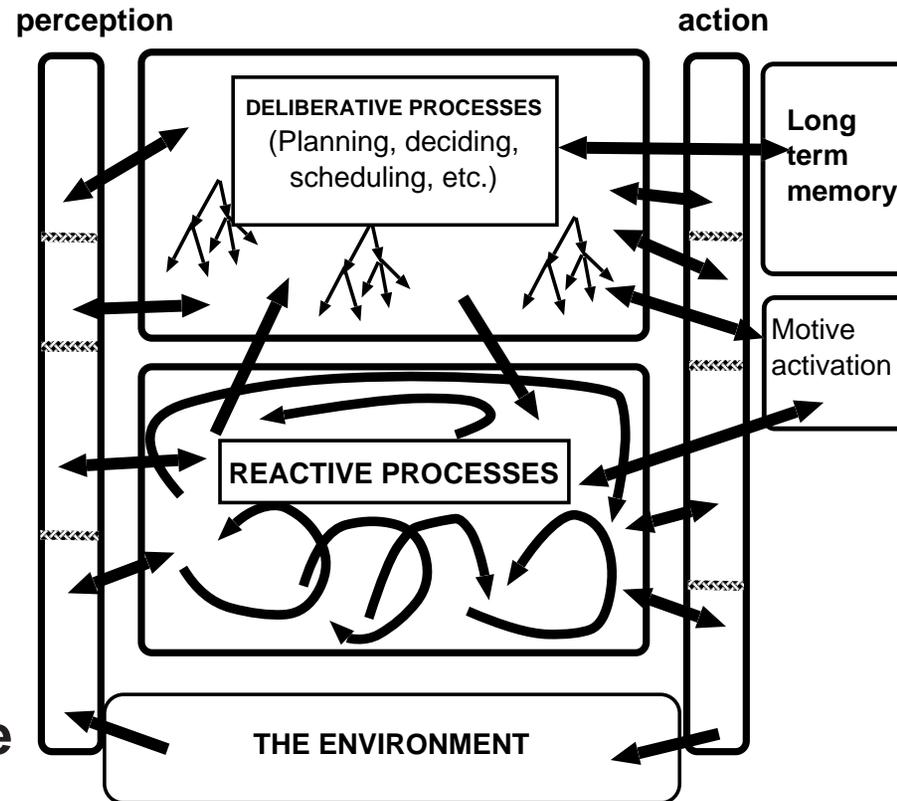
There may also be hierarchical action subsystems.

perception   action

ALARMS

REACTIVE PROCESSES

THE ENVIRONMENT
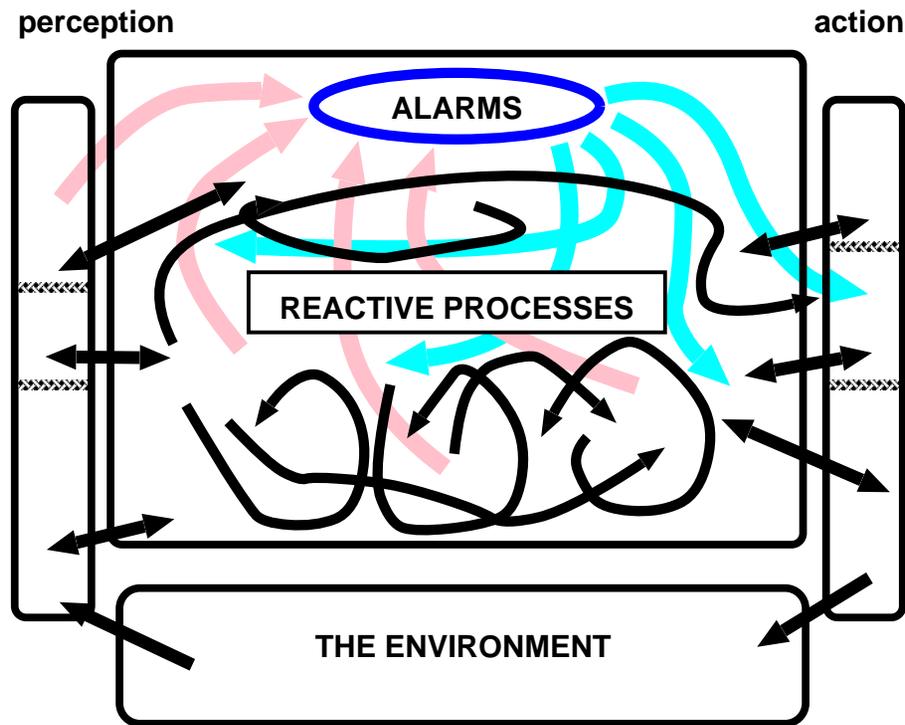
Towards deliberative systems

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

The requirements of a deliberative later could form a niche applying pressure for evolution of more abstract levels of perceptual processing, e.g. chunking perceptual inputs into forms useful for learning predictive associations, or for learning which actions do what.

I.e. perception evolves to support the needs of 'what if' reasoning mechanisms. (Many variants are possible.)

perception   action

DELIBERATIVE PROCESSES
(Planning, deciding, scheduling, etc.)

Long term memory

Motive activation

REACTIVE PROCESSES

THE ENVIRONMENT
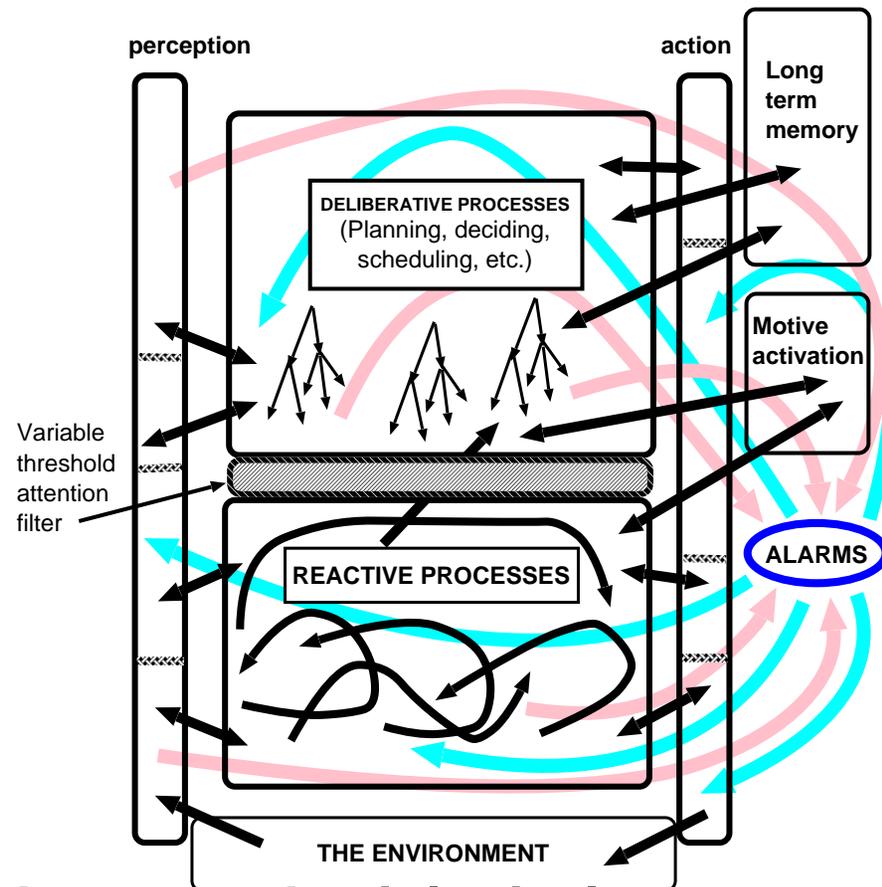
Alarm mechanism (global interrupt/override):

- Allows rapid redirection of the whole system
- sudden dangers
- sudden opportunities

- Freezing
- Fighting, attacking
- Feeding (pouncing)
- General arousal and alertness (attending, vigilance)
- Fleeing
- Mating
- More specific trained and innate automatic responses

What Damasio and Picard call "Primary Emotions" seem to be certain states generated in reactive mechanisms via global alarm systems.

Reactive and deliberative layers with alarms

Deliberative mechanisms come in various forms. The most sophisticated ones have complex architectural requirements, indicated only sketchily above.
What Damasio and Picard call "Secondary Emotions" seem to be reactions triggered by central cognitive processes in a deliberative mechanism.

perception

action

Long term memory

DELIBERATIVE PROCESSES
(Planning, deciding, scheduling, etc.)

Motive activation

Variable threshold attention filter

ALARMS

REACTIVE PROCESSES

THE ENVIRONMENT

Note: Whether these involve the same physiological responses as primary emotions in humans and other animals is an empirical question. There is no *theoretical* reason why they should *always* do so. Humans seem to vary in this respect.

H-COGAFF: A human-like architecture.

An instance of the CogAff schema using all the components.

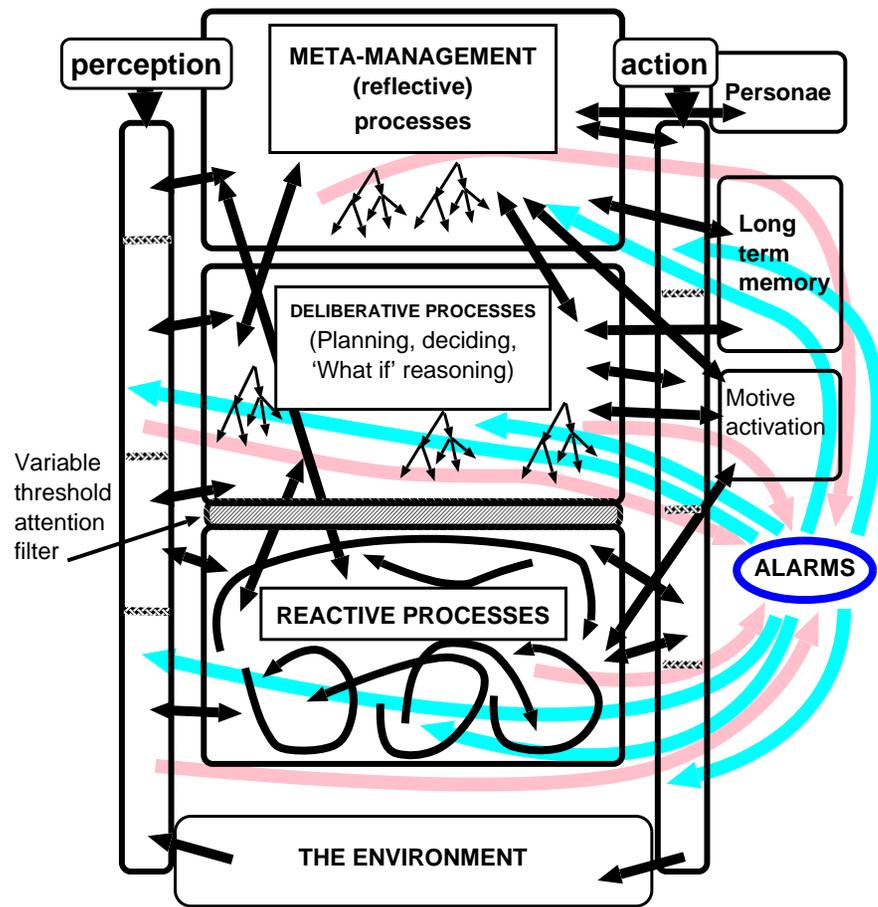The diagram is very impressionistic, not a precise "blue-print".
Described in more detail in papers in the Cogaff directory:
http://www.cs.bham.ac.uk/ research/cogaff/

Probably includes several alarm mechanisms. (Brain stem, limbic system, blinking reflexes, ...???)

Attention filter is needed to protect resource-limited deliberative and meta-management systems from relatively unimportant interrupts from reactive and alarm mechanisms.

But no filter is perfect, and some emotional states come from low importance interrupts given "high insistence" by stupid alarm systems: a type of "perturbance" (tertiary emotion).

perception

action

META-MANAGEMENT (reflective) processes

Personae

Long term memory

DELIBERATIVE PROCESSES (Planning, deciding, 'What if' reasoning)

Motive activation

Variable threshold attention filter

ALARMS

REACTIVE PROCESSES

THE ENVIRONMENT

Tertiary emotions

(Called "perturbances" in older Cogaff project papers.)

- **Involve interruption and diversion of thought processes.**

  I.e. the meta-management layer does not have complete control.

- **Question: Is it essential that all sorts of emotions have physiological effects outside the brain, e.g. as suggested by William James?**

  No: which do and which do not is an empirical question, and there may be considerable individual differences.

- **An organism that does not have meta-management cannot control attention, etc. and therefore cannot LOSE that sort of control, and therefore cannot have tertiary emotions.**

- **It does NOT follow that tertiary emotions are required for intelligent control.**

  (Damasio's non-sequitur.)

Different architectural layers support different sorts of emotions

→ **We can use the layers to define *architecture-based ontologies for different sorts of minds*.**

→ ***Describing different animals will require using different mental ontologies***

→ **Humans at different stages of development will instantiate different mental ontologies.**

Some notes:

➔ **Different aspects of love, hate, jealousy, pride, ambition, embarrassment, grief, infatuation can be found in all three categories of emotions.**

➔ **Remember that these are not *static* states but *developing* processes, with very varied aetiology. Different patterns of growth and decay correspond to different sorts of emotions.**

➔ **We don't necessarily already have names for all the significantly different cases**

➔ **Not all emotions are necessarily useful. Some can be seriously dysfunctional.**

➔ **Moods are global control states often confused with emotions. Attitudes (e.g. love of one's country) are specific cognitive states often confused with emotions.**

Socially important human emotions

→ **These involve rich concepts and knowledge and high level control mechanisms (architectures).**

→ **Some emotions use categories for self-description and self-evaluation that are absorbed from a culture.**

→ **Some socially important processes involve switching between different personalities in different social contexts.**

Example: longing for someone or something:

- **Semantics: To long for something you need to know of its existence, its remoteness, and the possibility of being together again.**

- **Control: One who has deep longing for X does not merely occasionally think it would be wonderful to be with X. In deep longing thoughts are often *uncontrollably* drawn to X. Moreover, such longing may impact on various kinds of high level decision making as well as the focus of attention.**

  Physiological processes (outside the brain) may or may not be involved. Their importance is over-stressed by some experimental psychologists.

Brains support consciousness? How?

What's consciousness?

People assume consciousness is one thing.

Then they ask questions like:

- **which animals have IT?**

- **how did IT evolve?**

- **what is ITS function?**

- **could machines have IT?**

- **which bits of the brain produce IT?**

If there's no "IT" the questions make no sense.

→ What we call "consciousness" is a large ill-defined COLLECTION of capabilities.

→ Not just ONE thing.

→ THEY can be present or absent in different combinations, in different animals, in people at different stages of development or after brain damage.
Also in different machines.

→ No pre-ordained subset of that set of capabilities is THE subset required for consciousness.

→ I.E. "CONSCIOUSNESS" IS A VERY VAGUE "CLUSTER CONCEPT". (Like "emotion")

→ People think they know what IT is from experience.
Before Einstein people thought they knew what simultaneity was from experience. We can unintentionally fool ourselves.

Varieties of consciousness

By exploring varieties of awareness of the environment and varieties of self-awareness made possible by different architectures we can distinguish different varieties of consciousness.

➜ **Microbe consciousness**

➜ **Flea consciousness**

➜ **Frog consciousness**

➜ **Eagle consiousness**

➜ **Chimp consciousness**

➜ **Infant (human) consciousness**

➜ **Adult consciousness**

➜ **Varities of drug-modified consciousness**

See talk 9 here http://www.cs.bham.ac.uk/˜axs/misc/talks/ (on varieties of consciousness.)

Qualia and meta-management.

➜ **The third layer enables**
  – SELF-MONITORING, SELF-EVALUATION
    AND
  – SELF-CONTROL
➜ AND THEREFORE ALSO LOSS OF CONTROL (TERTIARY EMOTIONS: PERTURBANCES)
➜ **and qualia**
  arising out of internal self-monitoring capabilities.

Sensory qualia depend on:

➜ **Existence of structured internal intermediate representations in perceptual mechanisms**

➜ **Ability of meta-management systems to attend to (inspect, analyse, compare, describe) the contents of internal representations, instead of the external scene.**

Robots with similar meta-management capabilities are likely to invent philosophical problems about qualia – and may wonder whether humans have them.

What we still don't know

➜ **The above ideas could lead us to significant progress despite many technical and conceptual difficulties.**

➜ **However our understanding of information processing architectures is still very rudimentary.**

   We have learnt a lot through work in software engineering, control theory, AI and other fields, but it is still very likely that we are nowhere near a full grasp of the richness and variety of architectures already produced by biological evolution and perhaps capable of being produced artificially in the distant future.

➜ **Some humility is required.**

Examples of ignorance:

→ **(a) Our general notion of "computation" has nothing to do with Turing machines: instead it is a product of two old strands of engineering development (i) control systems of many kinds (including looms, musical boxes, card-sorters) (ii) systems that operate on abstract entities, such as numbers and census information.**

There are many attempts in progress to produce new sorts of information processing machines including quantum computers, biologically inspired neural nets, DNA computers — yet we cannot know what new sorts of machines will prove useful in future decades or centuries.

More ignorance:

→ **(b) Many people have observed that human intelligence often involves the use of visual/spatial forms of thinking and reasoning even in connection with non-visual tasks (e.g. transfinite set theory). But nobody has the faintest idea how human visual systems represent and manipulate visual information in such a way as to facilitate all this.**

Some thoughts on what needs to be explained (not how to do it) are in talks 7 and 8 here:

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

E.g. what is it to see a blank surface? Compare a young child and a Picasso looking at the same surface: what's the difference? We need to find new ways to unravel the phenomenology of visual experience. (We can build on things like the discussion of "seeing as" in Wittgenstein's *Philosophical Investigations* part II section xi.)

More ignorance:

➜ **(c) There are many aspects of individual human development and learning that we do not begin to understand. What sort of architecture does a new-born infant have, and how does it bootstrap itself to the architecture of a toddler, then a youth then an adult. How does it enrich its innate ontology to think about culture-specific things such as infinite sets, fairy tales, religious dogmas, politics, television programmes, computers, theories of mind, quantum physics... ?**

(Perhaps we can begin to get some clues by trying to understand differences between precocial and altricial species.)

More ignorance:

➤**(d) Part of our problem in trying to understand products of evolution is that we are often "ontologically blind" to certain important features of those systems.**

This is a general problem at the frontiers of science: someone with restricted conceptions of motion may perceive and think about how fast objects move, but never notice that besides a velocity objects can have an instantaneous acceleration. Likewise, a researcher who thinks the function of visual systems is to provide information about geometrical shapes, motion, distances, and colours (Marr 1982) may never notice situations where vision provides information about abstract relationships between relationships (Evans 1968), information about affordances (Gibson), e.g. graspability, obstruction, danger, opportunity, or information about causal relationships, e.g. when a rope restricts the movement of a stick to which it is tied (Kohler 1927).

It is very likely that the questions asked in this paper are pale shadows of the questions we shall have answered a few centuries hence.

Learning to think outside your current ontology may seem to be impossible by definition: yet it is the basis of all major scientific advances (see chapter 2 of Sloman 1978 – Online at my web site). Maybe we need new ways of training scientists to do that.

More ignorance:

➜ **(e) There are many unsolved problems of representation, including how to represent modal notions required for perceiving and understanding affordances. These include concepts of possibility, impossibility, causation and counterfactual conditionals. The formal systems developed by logicians (modal logics) do not seem to be well suited to explaining their role in natural intelligence or intelligent robots of the future. For some preliminary thoughts about this see (Sloman 1996).**

➜ **(f) Besides *architecture-based* concepts there are also *architecture-driven* concepts, i.e. concepts whose best explanation may turn out to rely on a description of the sort of architecture that deploys them (as opposed to what they refer to). E.g. the concepts of number grasped by a child learning mathematics may depend crucially on ways in which the child's architecture supports various operations like counting, and meta-observations of properties of the counting process.** (See chapter 8 of Sloman 1978)

There is still a huge amount of research (including clarification of the questions) to be done on this.

More ignorance (aesthetics and jokes):

→ **(g) There are many commonplace phenomena of mind that researchers in AI, psychology and cognitive science so far do not seem to have any grip on: experiencing aesthetic qualities, including finding something beautiful, ugly or funny.**

People have designed programs that can generate very high quality music (David Cope), or extraordinarily good pictures (Harold Cohen), or moderately funny jokes (Kim Binsted). But in general those programs are not very good at telling which of the things they produce are good and which bad or why, although in more advanced systems there are at least the beginnings of such distinctions in the ways in which the programs make some of their choices.

Even if a program could distinguish with great accuracy the things most humans would find beautiful or funny that would still not mean that the program was anywhere near having the experience of finding anything beautiful or funny itself. Neither do we know what would be required to make that possible.

Many attempts have been made to analyse such experiences (e.g. by Koestler, Freud and others) but I don't find any of them convincing. A convincing explanation would have to be part of a larger explanation of what it is to have experiences and to evaluate them in different ways. Maybe thinking about architectures will enable us to make some progress. Maybe not.

Why did the robot cross the road?

?????

Why did the robot cross the road?

> ## TO FIND OUT WHY
>
> ## THE JOKE ABOUT THE CHICKEN WAS FUNNY

Most attempts to explain what it is to find something funny talk about how humour might have evolved, or what the social usefulness of humour is, or what kinds of circumstances produce it, or what the consequences of finding something funny are, e.g. the behaviour it produces.

What we really need is an understanding of how the ability to experience things as funny emerged from mechanisms that evolved because of their biological usefulness.

A theory that simply postulates a "humour box" in the architecture will have no explanatory power.

Likewise theories that postulate "emotion boxes" in the architecture have no explanatory power. By contrast we explain the need for *alarm* mechanisms in terms that a designer/engineer can understand, then show how certain processes with features we recognize as emotions sometimes emerge from their operation.

???

→ **(h) Actually, we don't know what we don't know....**

Acknowledgements

This research is supported by a grant from the Leverhulme Trust.

I have had much help and useful criticism from colleagues at Birmingham and elsewhere, especially Luc Beaudoin, Ian Wright, Brian Logan, Matthias Scheutz and Ron Chrisley. There is considerable overlap with ideas about architectures in the work of Marvin Minsky, e.g. in *The Society of Mind* and in his draft book *The Emotion Machine* available on his web site: http://web.media.mit.edu/˜minsky/

The Birmingham Cognition and Affect Project

PAPERS (mostly postscript and PDF):

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

(References to other work can be found in papers in this directory)

TOOLS:

http://www.cs.bham.ac.uk/research/poplog/freepoplog.html

(Including the SIM_AGENT toolkit)

SLIDES FOR TALKS (Including IJCAI01 philosophy of AI tutorial with Matthias Scheutz):

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

Free online book: The Computer Revolution in Philosophy (1978)

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