Presented at Gatsby Computational Neuroscience Unit

Varieties of Affect and the CogAff Architecture Schema

AARON SLOMAN

http://www.cs.bham.ac.uk/~axs/
A.Sloman@cs.bham.ac.uk

School of Computer Science
The University of Birmingham

This and other slide presentations can be found at
http://www.cs.bham.ac.uk/~axs/misc/talks/

Related papers can be found at
http://www.cs.bham.ac.uk/research/cogaff/

Acknowledgements

Ideas presented here were developed in collaboration with

Steve Allen, Luc Beaudoin, Ron Chrisley,
Catriona Kennedy, Brian Logan,
Riccardo Poli, Matthias Scheutz,
Ian Wright

and others in the Cognition and Affect Project

School of Computer Science
The University of Birmingham

http://www.cs.bham.ac.uk/˜axs/cogaff.html

Closely related work can be found at Marvin Minsky's web site, including his draft book The Emotion Machine, available at http://www.media.mit.edu/˜minsky/

This work is supported by a grant from the Leverhulme Trust

http://www.leverhulme.org.uk

"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

Abstract

In the last decade and a half, there has been a steadily growing amount of work on affect in general and emotion in particular, in empirical psychology, cognitive science and AI, both for scientific purposes and for the purpose of designing synthetic characters, e.g. in games and entertainments.

Such work understandably starts from concepts of ordinary language (e.g. “emotion”, “feeling”, “mood”, etc.). However, these concepts can be deceptive: the words appear to have clear meanings but are used in very imprecise and systematically ambiguous ways. This is often because people use explicit or implicit pre-scientific theories about mental states and process which are incomplete or vague. Some of the confusion arises because different thinkers address different subsets of the phenomena.

More sophisticated theories can provide a basis for deeper and more precise concepts, as has happened in physics and chemistry following the development of new theories of the architecture of matter which led to revisions of our previous concepts of various kinds of substances and various kinds of processes involving those substances.

In the Cognition and Affect project we have been exploring the benefits of developing architecture-based concepts of mind. We start by defining a space of architectures generated by the CogAff architecture schema, which covers a variety of information-processing architectures, including, we think, architectures for insects, many kinds of animals, humans at different stages of development, and possible future robots.

In this framework we can produce specifications of architectures for complete agents (of various kinds) and then find out what sorts of states and processes are supported by those architectures. Thus for each type of architecture there is a collection of “mental concepts” relevant to organisms or machines that have that sort of architecture.

Thus we investigate a space of architectures linked to a space of possible types of minds, and for some of those minds we find analogues of familiar human concepts, including, for example, ”emotion”, ”consciousness”, ”motivation”, ”learning”, ”understanding”, etc.

Abstract continued

We have identified a special type of architecture H-Cogaff, a particularly rich instance of the CogAff architecture schema, conjectured as a model of normal adult human minds. The architecture-based concepts that H-Cogaff supports provide a framework for defining with greater precision than previously a host of mental concepts, including affective concepts, such as ”emotion”, ”attitude”, ”mood”, ”pleasure” etc. These map more or less loosely onto various pre-theoretical versions of those concepts.

For instance H-Cogaff allows us to define at least three distinct varieties of emotions; primary, secondary and tertiary emotions, involving different layers of the architecture which we believe evolved at different times. We can also distinguish different kinds of learning, different forms of perception, different sorts of control of behaviour, all supported within the same architecture.

A different architecture, supporting a different range of mental concepts might be appropriate for exploring affective states of other animals, for instance insects, reptiles, or other mammals. Human infants probably have a much reduced version of the architecture which includes self-bootstrapping mechanisms that lead to the adult form.

Various kinds of brain damage can be distinguished within the H-Cogaff architecture. We show that some popular arguments based on evidence from brain damage purporting to show that emotions are needed for intelligence are fallacious because they don’t allow for the possibility of common control mechanisms underlying both tertiary emotions and intelligent control of thought processes. Likewise we show that the widely discussed theory of William James which requires all emotions to involve experience of somatic states fails to take account of emotions that involve only loss of high level control of mental processes without anything like experience of bodily states.

We have software tools for building and exploring working models of these architectures, but so far model construction is at a very early stage.

PLAN FOR THE TALK

(Needs a week, or a month, actually)

1. **A problem:**
   how do our concepts of mind work? We don't know, but we think we know — our ideas about mind 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, when one feels a tusk, another an ear, another leg, etc.
3. **Recommendation:**
   a good way forward (depending on your motivation in studying emotions):
   study architectures for possible minds and architecture-based concepts of mind, and use them to generate architecture-based concepts.
4. **Vague Pointers to Evidence:**
   a growing collection of supporting evidence from common knowledge, psychology, psychiatry, neuroscience, ... e.g. the book by Barkley, mentioned later.

See also this talk on Varieties of Consciousness:

http://www.cs.bham.ac.uk/˜axs/misc/talks/#talk9

Different motivations for interest in implementable models of emotions

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

(ii) Improved interaction:
A desire to give machines which have to interact with humans an understanding of emotions as a requirement for some aspects of that task (Sloman 1992)

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

(iv) Education:
Using models of type (i), (ii), (iii) etc. in educational tools for trainee psychologists, therapists, etc.

(v) Therapy, counselling, etc.:
If we have a better understanding of the nature of the emotions and other affective states, and their architectural underpinnings, we may be better able to provide helpful therapy when needed.

The conceptual requirements for these objectives are different.

E.g. “believable” behaviour in constrained contexts could be the product of widely different models, including at one extreme very large, hand-coded lookup tables specifying what to do when.

But in the long run a deep and accurate model of the first type may be required for effectively achieving even the engineering goals of types (ii) and (iii)

For now we address only goal (i) (Science and philosophy, including conceptual analysis), while keeping an eye on the requirements for the others.

NOTE: I am not specially interested in *emotions* except as a special case of a wide range of phenomena that need to be accommodated in a theory of what minds are (an ontology for minds) and explanations of how they work.

Here is a very simple toy demo of type (iii)/(iv).

(Show sim_feelings demo
http://www.cs.bham.ac.uk/research/poplog/sim/teach/sim_feelings)

PROBLEM:

Do we understand what we mean by: "consciousness" "emotion" "intelligence" "mind" "learning" etc .... ???

We successfully use language to talk and think about mental states and processes. We understand this well enough for purposes of everyday communication. But we don't have self-knowledge about *what* we mean.

We *think* we do: so we propose definitions with great confidence.

However, different thinkers produce WIDELY differing definitions, of "emotion", "mind", "intelligence", "understanding", "creativity", "consciousness", ...

OUR CONFIDENCE IN OUR DEFINITIONS IS MISPLACED

This does not matter much for everyday interactions, writing novels, etc.

But for scientific purposes we MUST diagnose the causes of the confusion, and find better concepts and theories.

Try getting friends and acquaintances to write down a definition of "emotion". Then compare your definitions.

NB Include people from different disciplines and professions.

EXAMPLES OF CONFUSIONS

Many thinkers disagree on answers to these questions:

- **Is surprise an emotion?**

  (Some people say “always”, others say “only in certain cases”.)
- **If you love your country, is that an emotion, or an attitude?**

  What if your love for your country is far from your present thoughts?
- **Can you have an emotion without being aware of it?**

  E.g. jealousy, infatuation
- **Does an emotion have to have some externally observable/measurable physiological manifestation?**
- **Can a fly feel pain, or have emotions?**
- **Is there a stage at which a human foetus becomes able to have emotions?** (E.g. able to worry about how the birth will go?)
- **Could a disembodied mathematician have emotions?**

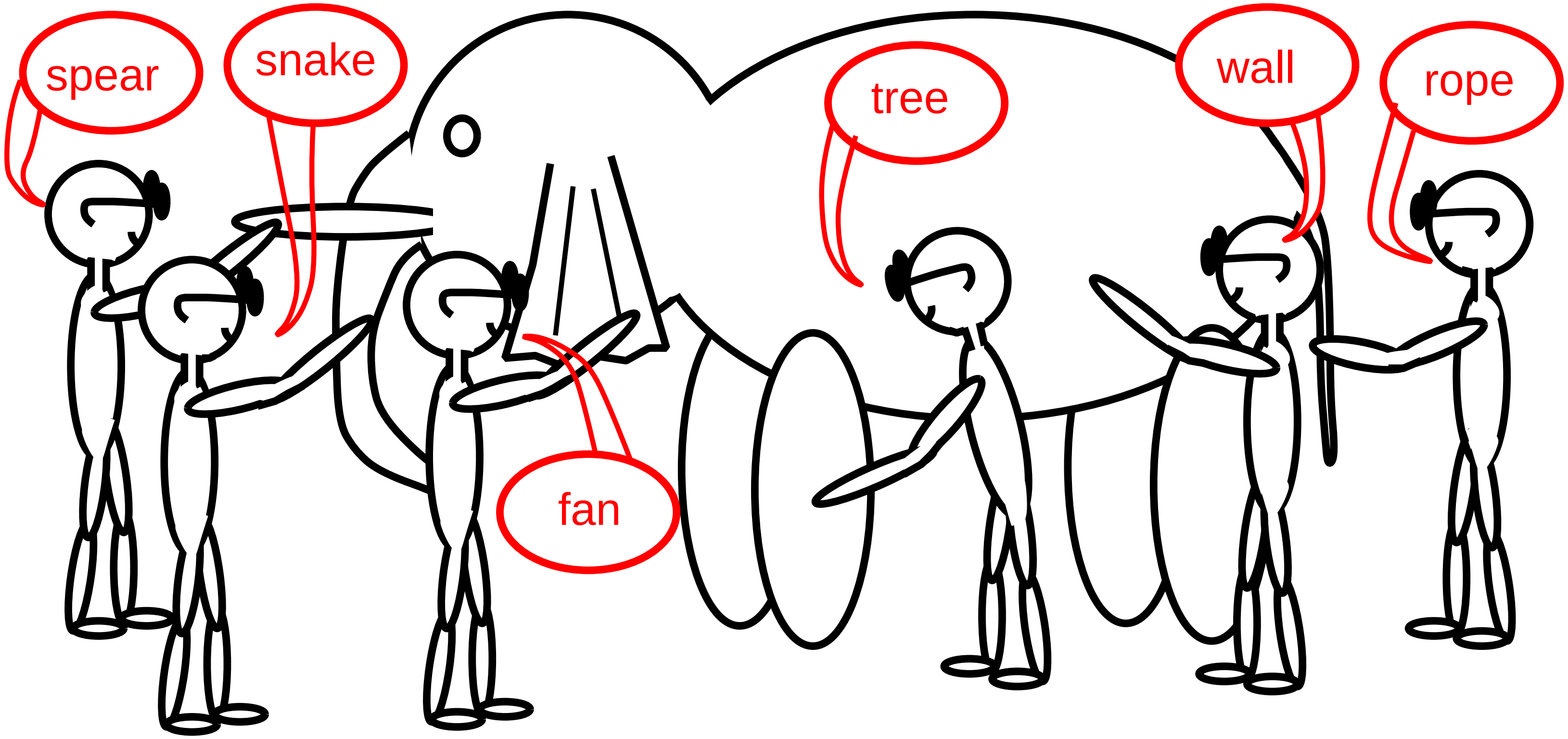
  (E.g. feel disappointment at finding a flaw in a proof?)

There is no consensus about what emotions are

IT’S WORSE THAN SIX BLIND MEN DESCRIBING AN ELEPHANT!

What is an Emotion :: 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



Can anyone see the whole reality?

There are many prejudices that confuse these investigations:

E.g. Some people assume:

- **That minds must be "embodied" and emotions must involve bodies**
  (Ignoring the possibility of passionate disembodied mathematicians.)
- **That emotions are required for intelligence.**
  (Damasio's non-sequitur (1994))
- **That emotions evolved because they are useful**
  (Maybe *some* did. It does not follow that *all* are useful.
  Many may be side-effects of *other* useful mechanisms.)
- **That you can't have an emotion without being conscious of it.**
  (What about the person who is obviously infatuated or jealous but unaware of the fact?)
- **That GOFAI methods have failed, and must be abandoned.**
  (Just because the silly predictions of some early AI researchers did not come true!)
- **That it would be a "bad thing" if robots could have emotions.**
  (Could they possibly do worse things than humans do to other humans?)

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

CONFUSIONS ABOUT “EMOTION”

Many different definitions:

in psychology, philosophy, neuroscience, ethology ...

and many variants within each discipline

Some will be illustrated later.

PARTIAL DIAGNOSIS:

Different theorists concentrate on different phenomena.

We need a theory that encompasses all the phenomena.

Remember the six blind men trying to describe an elephant?

Some of them are feeling a hippopotamus.

Let’s rephrase the questions:

1. What are the architectural requirements for various kinds of mental states and processes in humans and other animals?

   And in various kinds of more or less intelligent machines.

2. What sorts of states and processes can each architecture support?

   And how do our pre-scientific mental concepts map onto those kinds of states: compare the relation between pre-scientific concepts of kinds of physical stuff and concepts based on current theories of the architecture of matter.

After exploring architectures and the states and processes they support

INSTEAD OF SITTING IN OUR ARMCHAIRS AND INVENTING DEFINITIONS, WE CAN:

- Collect examples of many types of real (and theoretically possible) phenomena,
  (in humans of many types (young, old, normal, brain-damaged, etc.) and animals of many types).
- Then try to build a theory which explains them all ....
- Subject to constraints from neuroscience, psychology, biological evolution, feasibility, tractability, etc.
  (E.g. R. A. Barkley, *ADHD and the nature of self-control*, 1997.)
- We can then define concepts corresponding to possible states and processes that can occur in the various architectures.

  The new architecture-based concepts can then extend and refine, rather than eliminate and replace, our old concepts.

NOTE: That's what has happened in the physical sciences.

A GOOD EXPLANATORY THEORY MUST ALLOW FOR VARIATION

(different clusters of capabilities):

There are different sorts of variation to be taken into account:

- Across species,
- Across individuals within a species,
- Within an individual during normal development
- After brain damage
- Across planets (grieving, infatuated, Martians?)
- Across the natural/artificial divide.

THE ARCHITECTURES WILL VARY:
so the states and processes that can occur will vary
and the concepts that are applicable will vary.

WHY STUDY VARIETIES OF ARCHITECTURES?

- **As philosophers: we want to understand the space of possibilities, and their implications**
- **As scientists we are likely to miss things if our search is too focused.**
- **As engineers we want to know what is achievable and how.**
- **You don't really understand an architecture unless you understand its advantages and disadvantages compared with neighbours in design space.**

  (Trade-offs in different contexts, etc.)
- **And in any case there are very different cases in nature.**

  Even among humans:

  Infants, children, autistics, people with ADHD, altzheimers patients, etc.

A TIP:
WHEN DOING CONCEPTUAL ANALYSIS

If someone puts forward a definition of mental state or process X, I find that asking:

> **“Does this also apply to X in flies, rats, chimpanzees, newborn infants?”**

often reveals that the definition was not based on sufficient thought.

A Control-based conception of emotion

What is there in common between

- a crawling woodlouse that rapidly curls up if suddenly tapped with a pencil,
- a fly on the table that rapidly flies off when a swatter approaches,
- a fox squealing and struggling to escape from the trap that has clamped its leg,
- a child suddenly terrified by a large object rushing towards it,
- a person who is startled by a moving shadow when walking in a dark passageway,
- a rejected lover unable to put the humiliation out of mind
- a mathematician upset on realising that a proof of a hard theorem is fallacious,
- a grieving parent, suddenly remembering the lost child while in the middle of some important task?

Proposed Answer:

in all cases there are at least two sub-systems at work in the organism, and one of them, a specialised sub-system, somehow interrupts or suppresses or changes the behaviour of others, producing some alteration in (relatively) global (internal or external) behaviour of the system.

Some people would wish to emphasise a role for *evaluation*: the interruption is based at least in part on an assessment of the situation as good or bad. Is a fly capable of evaluation?

All this is very vague

We can make it more precise by:

– spelling out different kinds of information-processing control architectures in which such things (e.g. global interrupts or modulations of processing) can occur

– showing how different varieties of states with these general features can arise in different architectures.

Different sorts of emotions (and other affective states) arise out of different sorts of:

– Interacting sub-systems

– Ways one can interrupt or modulate another

– Functional roles and side-effects

Emotions are a subclass of “affective” states

Affective states are of many kinds. They include not only what we ordinarily call emotions but also states involving desires, pleasures, pains, goals, values, ideals, attitudes, preferences, and moods.

The general notion of “affective state” is very hard to define but very roughly it involves using some kind of information that is compared (explicitly or implicitly) against what is happening, sensed either internally or externally.

– **When there’s a discrepancy some action is taken, or tends to be taken to remove the discrepancy by acting on the sensed thing: affective states involve a *disposition* to change reality in some way to reduce a mismatch.**

– **In contrast, if the information is part of a percept or a belief, then detecting a discrepancy tends to produce a change in the stored “reference” information.**

This is related to what some philosophers have called the difference in “direction of fit” between beliefs and desires. We generalise this to a larger class of information states. There are some which are neither belief-like nor desire-like, insofar as they merely represent possibilities which are being thought about or considered as options.

There is a more primitive type of control state which does not use any sort of description or representation that can be compared with reality, but merely generates action, or has a disposition to produce action (including resisting change). Many innate behaviours are like that. Are these “affective” states? There’s no *right* definition of such a vague notion.

Affect and architecture

As with emotions, which affective states are possible in an organism or machine will depend on the information-processing architecture of the whole system.

We'll consider three architectural layers:

reactive,

deliberative

and

meta-management,

and the different sorts of emotions that can be associated with them.

Which human-like states and processes?

There are AT LEAST THREE different classes of mental phenomena commonly referred to as "emotions".

- **Primary emotions**

  (Evolutionarily oldest – depend only on reactive mechanisms)
- **Secondary emotions**

  (Deliberative mechanisms generating these evolved later)
- **Tertiary emotions**

  (Newest and rarest: involve disruption of meta-management, e.g. loss of control of attention. These are usually not distinguished from secondary emotions)

NOTES:

– These rather vaguely defined categories, described below, are re-defined in terms of the information-processing architectures (virtual machine architectures) that make them possible. An animal without deliberative mechanisms cannot have secondary emotions.

– This is only an introduction to the diversity of types of emotions (and affect) and the list of types will probably need to be extended after further analysis.

Which kinds of emotions are of most interest in human relations (e.g. which kinds are referred to most in plays, novels, poems, garden fence gossip – and hardest to study in laboratories!)?

Primary emotions:

(Discussed by Damasio, Picard, Goleman, etc.)

Examples of primary emotions familiar in humans

- Being startled by a loud noise,
- Being frozen in terror as boulder crashes towards you,
- Being nauseated by a horrible smell

THESE REQUIRE ONLY EVOLUTIONARILY OLD REACTIVE MECHANISMS.

In primary emotions, sensor states and/or internal reactive states trigger a fast but stupid reactive "alarm" mechanism that produces global changes in motors and internal reactive states.

– Simple versions occur even in insects: when flee, fight, feed, freeze, or mate responses override other processes. (The five Fs!)

– In humans these primary emotions often have sophisticated accompaniments that cannot occur in most other animals capable of having primary emotions.

– E.g. when we are aware of having them we are using meta-management mechanisms that are not needed for primary emotions.

Often the primary emotion will immediately trigger some other kind, e.g. apprehension, a secondary or tertiary emotion.
Thus pure primary emotions are rare in humans.

Secondary emotions

Examples — You are:

- Afraid the bridge you are crossing may give way,
- Relieved that you got to the far side safely,
- Afraid the bridge your child is crossing may give way,
- Worried about what to say during your interview,
- Undecided whether to cancel your holiday in ...
- Enjoying the prospect of success in your endeavour,

Secondary emotions are triggered by events in a deliberative sub-system. Some of these are triggered by thinking about what might happen, what might have happened, what did not happen, etc., unlike primary emotions which are triggered only by actual occurrences.

So secondary emotions require deliberative capabilities with ‘what if’, i.e. counterfactual, representational and reasoning capabilities. These are very subtle and complex requirements.

Probably very few animals: Chimps? Bonobos? Gorillas?
Perhaps some other mammals?

(Damasio, Picard, Goleman, etc. use a more comprehensive definition, since they don’t distinguish deliberative from meta-management mechanisms.)

Tertiary emotions:

(Previously called “perturbances” by Sloman, Beaudoin, Wright)

Examples — You are:

- Infatuated with someone you met recently,
- Overwhelmed with grief,
- Riddled with guilt about betraying a friend,
- Full of excited anticipation of a loved one’s return,
- Full of longing for your mother,
- Basking in a warm glow of pride after winning an election.
- Obsessed with jealousy about a colleague’s success,

These involve *disruption* of high level *self* monitoring and control mechanisms. I.e. there is (actual or dispositional) loss of control of thought processes. Thus they cannot occur in animals and machines that are incapable of having such control.

An architecture including meta-management capabilities is required for tertiary emotions.

People who lump deliberative and meta-management mechanisms together under “executive functions” are unlikely to separate tertiary emotions from secondary emotions.

Emotions and architectures

In all of the categories (primary, secondary, tertiary emotions) there is a subsystem that produces some relatively *global* changes in the rest of the system, or in much of it.

They differ in

- **What kind of subsystem does the disrupting**
- **Where the information comes from that triggers the disrupting (e.g. does it come from a deliberative layer, or only sensors and internal states of a reactive layer?)**
- **Which parts of the system are disrupted, e.g. is there externally visible behaviour or only internal disruption? Which internal parts?**
- **Also there are differences in kind of semantic content, time scale, what can and cannot suppress the disruption, whether learning is involved, etc.**

In humans, primary, secondary and tertiary emotions, are not mutually exclusive

All three kinds of emotional processes can coexist in complex situations.

As a result of this, the emotions labelled in ordinary language, e.g. "fear", "anger", "relief", "distress", cannot simply be classified as primary, secondary or tertiary. Often they are a mixture.

People involved in long and tiring adventure trips often describe multiple emotions at the end. E.g. they may be simultaneously:

- **Glad to have succeeded in their aims**
- **Regretful at not having done better**
- **Sad that the trip is over**
- **Relieved that some threat did not materialise (e.g. running out of fuel).**
- **Glad to see their families again, etc.**
- **Hoping to be selected for their national team,**
- **Desperately longing for a good meal,**
- **Worried about an injury incurred on the trip, ...**

etc.

Different architectural underpinnings are required for different categories of emotions

- **Primary emotions**:
  Require sensors linked to fast reactive mechanisms that can sometimes trigger rapid global signal patterns sent to motors and other sub-systems.
- **Secondary emotions** (central and peripheral):
  Require signals from deliberative mechanisms to fast reactive mechanisms that can under certain conditions trigger rapid global reactions.
- **Tertiary emotions** (with and without peripheral effects):
  Presuppose self-monitoring self-controlling meta-management systems that can be disrupted or modulated by other sub-processes.

All have many variants

Finer distinctions can be made when we understand the underlying architecture better.

E.g.

- “Purely central” *vs* “partly peripheral” secondary emotions.
- Second-order emotions (being ashamed of feeling jealous).
- Deliberately induced emotions (teacher who – reluctantly – allows himself to get angry to achieve control of a difficult class)
- Emotions that involve constant activity (plotting, fretting, fuming, ranting).
- Emotions that vary in intensity over time.
- Long term, mostly dormant, emotions, e.g. jealousy, grief. (Often ignored.)

COROLLARY OF ARCHITECTURE-BASED APPROACH

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 – especially if they are not state-free.
- Likewise the category of 'deliberative' systems divides into sub-categories when we study different architectures. (Many other examples.)

Demonstrations available:

- 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.

OUR LACK OF SELF-KNOWLEDGE IS PERVASIVE

Do you know how you recognise a face in different lighting or when seen from different angles?

Do not assume you know what you mean by “emotion”, “consciousness”, “pleasure”, “pain”,“mood”, etc.

COMPARE: we can use the syntax of English (French, Urdu, etc.) effortlessly when we speak and hear utterances in a language we know.

But we have very little insight into the grammar of our language, and how the semantics of complex sentences are composed from the semantics of components. We have very little insight into ANY of the workings of our own minds.

Both our syntax and our conceptual structures are opaque to us: finding out what they are and how they work requires scientific investigation.

We must collect evidence and we must build theories that are rich in explanatory power.

Most mental concepts do not have clear or simple definitions: they are partly indeterminate “cluster concepts”. See http://www.cs.bham.ac.uk/research/cogaff/ibm02

The CogAff project

The Cognition and Affect project provides a framework for designing architectures and exploring ways of basing concepts on them.

PLEASE JOIN US.

WE INVESTIGATE ARCHITECTURE-BASED CONCEPTS OF MIND

This involves

- **Exploring possible explanatory architectures**
- **Finding out which sorts of concepts are supported by different sorts of architectures (by analysis and by running programs where possible.)**
- **Trying to find which architectures provide good explanations for known types of animal minds, human and otherwise.**
- **Using that to decide which mental concepts are likely to be applicable to which organisms.**

Example:

Insects may be able to have (simple) primary emotions (as defined here) but not secondary or tertiary emotions (unless we are wrong about their information-processing architectures).

THERE ARE MANY OTHER APPROACHES TO THE STUDY OF EMOTIONS

Some inadequate approaches (1)

1. Definitions in terms of behaviour and behavioural dispositions.

These don't work because any collection of behaviours (and behavioural dispositions) can arise out of arbitrarily many different causal mechanisms.

Totally convincing "emotional" behaviours could be the product of an expert conman, or a powerful actor on a stage, or a huge Eliza-like program.

2. Definitions in terms of neural or physiological mechanisms.

But it is unsatisfactory to rule out apriori the possibility of alien species or robots being jealous, angry, excited, disappointed, etc.

Definitions that require production and detection of "somatic" states fall into this category. They rule out kinds of anxiety, obsession, grief whose manifestations are all mental.

(Of course mental states and processes in humans involve physiological processes in brains. But there's no need to *define* virtual machine states in terms of their implementation details.)

Some inadequate approaches (3)

4. Definitions that cover all affective states, e.g. desires, preferences, enjoyment, etc.

These may be useful for some purposes but disregard the possibility of wanting something and being completely unemotional about it (a “cold” desire). It is also possible to have certain pleasures, e.g. enjoying food or fresh air on coming out of a stuffy room, without being emotional in the ordinary sense of “emotional” – e.g. one may be perfectly calm and in total control.

Not every evaluation is an emotion.

5. Ostensive definitions based on “first person” experience.

These don’t work, though they seduce many scientists and philosophers.

Being able to recognize a subset of instances and non-instances, whether internal or external, does not require a full explicit understanding of the general principles involved.

Compare: thinking you have a grasp of the concept of simultaneity because you have first-hand “direct” experience of simultaneity.

Before Einstein, the hidden complexity of “X and Y happened at the same time” went unnoticed. Experiencing simultaneity gives the illusion of knowing what it is.

Support from a neuropsychiatrist

This remarkable book:

Russell A. Barkley, (1997)
***ADHD and the nature of self-control*,**
The Guildford Press, New York

(Partly inspired by writings of Jacob Bronowski)

has a detailed analysis of empirical evidence relevant to some of the mechanisms that we have assumed constitute meta-management.

Though he is a neuropsychiatrist (apparently a leading authority on ADHD) he has somehow learnt to think like a designer: a software engineer.

However, his architectural specification, like ours, needs further detail: being linked to empirical evidence and brain locations is not the same as being precisely specified.

Evolution gave us the “design stance”

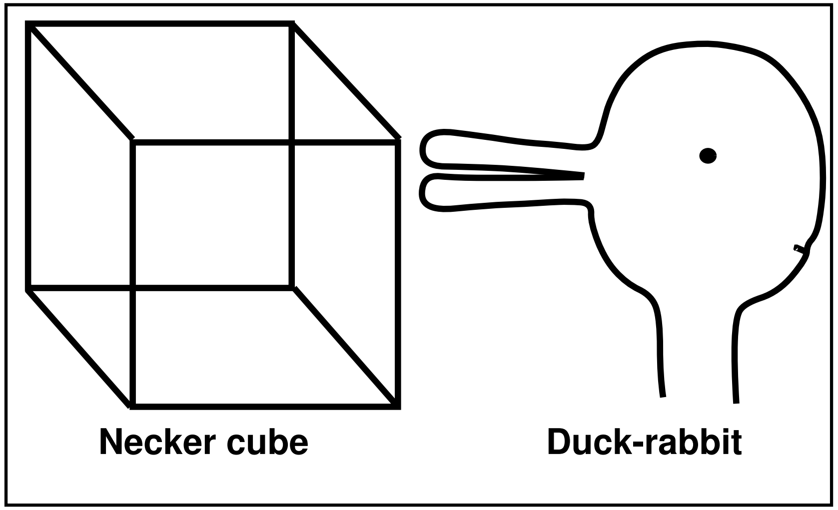
Evolution somehow developed species that use concepts of mind both for some of their own internal processing (meta-management, described later) and for dealing with other intelligent animals (predators, prey).

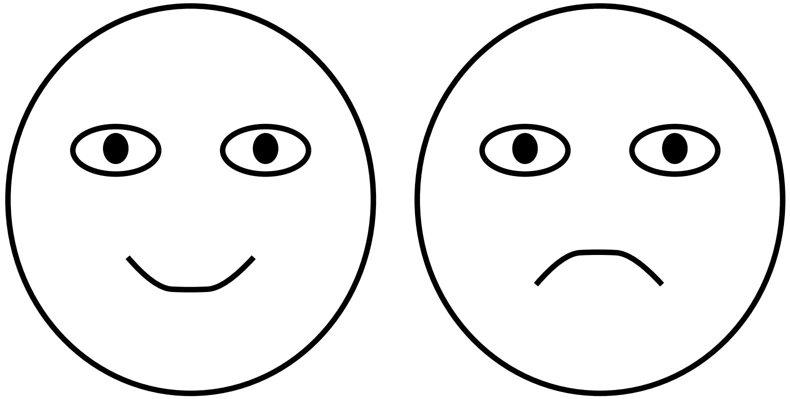
- I.e. an ontology that is useful for self-management may also be useful for reasoning about others, and vice-versa.
- Some aspects of the ontology may be architecture-based (as opposed to simply being related to perceived patterns of behaviour).
- E.g. a system with meta-management categorises its internal states in terms of percepts, desires, beliefs, intentions, attention, arising out of processes in different sub-systems.
- Thinking of others as having a similar architecture may be a powerful aid to dealing with others. (Contrast Dennett’s “intentional stance”.)

Conjectures:

- In humans (and some other animals?) perceptual mechanisms evolved to use such ontologies – e.g. seeing someone as looking happy, facing to the left.
- Our intuitive concepts of mind are inherently architecture-based: I.e. an implicit architecture-based theory of mind is (partly) innate.

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.

(Contrast Marr's views on vision, and much AI vision research.)

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...

Why would the ability to perceive mental states evolve?

Think about it:

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

- The shape and motion of X's body?

  OR
- Whether X is hungry?
- Whether X can see you?

Primitive implicit theories of mind probably evolved long before anyone was able to talk about theories of mind. Compare other intelligent primates.

(Evolution solved the "other minds" problem before there were any philosophers to notice the problem.)

Theories of mind produced by evolution

- An organism with some partial ability to monitor, describe, evaluate its own internal states needs an ontology for that purpose.
- The ontology need not be deep or accurate, as long as it works (compare the "naive physics" used by animals to interact with the physical environment).
- The same ontology might be capable of being used also to perceive, think about and interact with OTHER intelligent organisms.
- Using the same ontology is compatible with using quite different mechanism for recognising instances:
  - Using *internal* perceptual mechanisms for one's own mental states, processes, events
  - Using *external* perceptual mechanisms for the mental states, processes, events in other organisms.
  - Both *internal and external* perception depend on evolutionary changes both in the things being perceived and the mechanisms for perceiving them. (Compare Darwin on evolution of emotional expression.)

Primitive theories support primitive concepts

The theories of mind that evolved to meet biological necessities, are likely to be no more deep or accurate than primitive theories of matter that suffice for moving around in a physical world.

- **Both are adequate for their (pre-scientific) purposes.**
- **But both cause problems when used for more sophisticated purposes.**

Deeper, richer, more precise theories of the architecture of mind can provide a basis for more powerful sets of concepts of mind.

How can we improve our pre-theoretical concepts of mind?

We need to improve our ordinary concepts for the purposes of scientific understanding

The colloquial concepts work fine for normal purposes of communication, e.g. reporting another person's state as "asleep", or "angry".

> **But our pre-scientific concepts are based on a poor implicit theory of the architecture:**
>
> **They are usually good enough for everyday life, but not good enough for scientific explanation or deep modelling.**

The verbally specified alternatives used by psychologists to motivate their experiments are not much better: most of them are not trained engineers.

(Likewise most philosophers. And Penrose etc.)

Normally, only someone with software engineering expertise can think clearly about information-processing architectures.

Architecture-based concepts of matter

Our theory of the architecture of matter supports

- **our concepts of kinds of matter**
- **our concepts of kinds of processes that involve changes of matter**

Our concepts of kinds of matter, kinds of physical stuff, were extended and refined as we learnt about the architecture of matter:

- **E.g. the periodic table of elements was explained by the theory of the architecture of sub-atomic physics.**
- **Understanding how atoms can and can't combine generates a space of chemical concepts.**
- **There are concepts of types of *process* e.g. catalytic reaction, as well as concepts of types of state.**

Similarly we can generate architecture-based concepts of mental states and processes.

Architecture-based concepts

Many of our concepts are "generated" within an ontology that defines an architecture, e.g. for matter, for mental mechanisms, for social systems, for political systems, for a computer operating system.

We constantly use ontologies referring to virtual machines with complex architectures, even if we are unaware of doing so.

Science extends, corrects, and refines our theories of the underlying architectures.

- These are generally architectures for "virtual machines", not just physical machines.
- But they are all ultimately implemented in physical machines.

This leads to many philosophical problems about supervenience, implementation, etc.

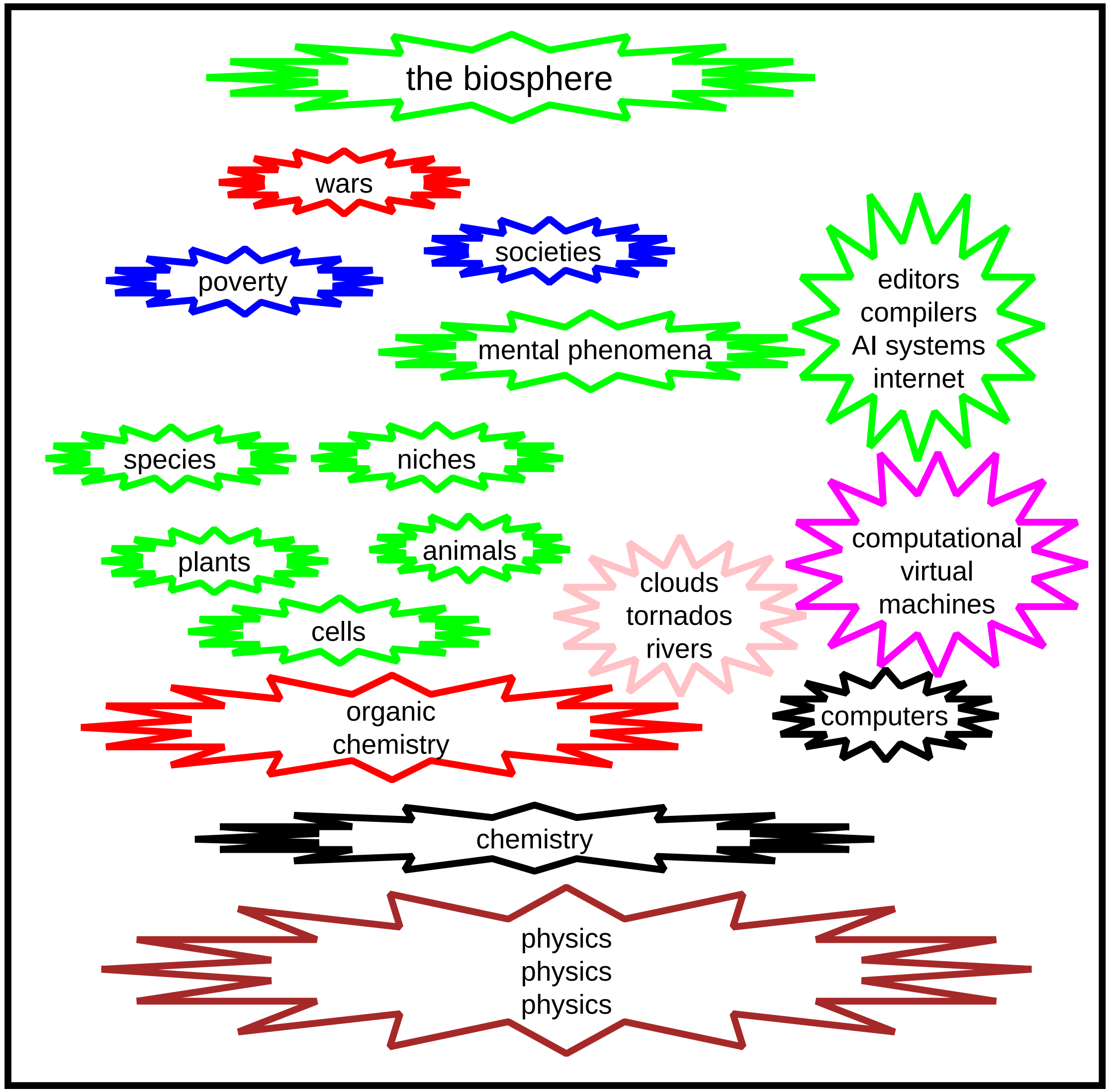
See the Cogaff web site, and our IJCAI 2001 tutorial on Philosophy of AI for more on this:

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

We barely understand the variety of virtual machine architectures

Emergent virtual machines are everywhere

How many levels of physics will there be in 500 years time?



An architecture supports a collection of concepts

When we understand the architecture of an operating system we can introduce new concepts referring to states and processes that can arise within that architecture:

E.g. the architecture of a computing system leads us to define:

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

and many more.

NOTES:

– Some of these concepts would not be applicable in an architecture that does not support concurrent multi-processing.

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

– It may use some if it does some self-monitoring!

There are many kinds of information-processing architectures, supporting many sets of concepts

Unlike physics:

- Physics has many levels, but there's still one physics subsuming them all (even if we don't know all the details yet).
- For minds there is not just ONE architecture but MANY, e.g. architectures for different animals.
- The different architectures support many different sorts of concepts of internal states.

- Flea minds (and emotions)
- Mouse minds (and emotions)
- Cat minds ...
- Chimp minds ...
- Human neonate minds ...
- Your mind ...

We understand only a tiny subset of the space of possible virtual machine architectures for organisms and machines.

Minds of different sorts need different VM architectures. E.g.

- **adult human minds, infant human minds,**
- **chimpanzee minds, rat minds, bat minds,**
- **flea minds,**
- **damaged or diseased minds ....**

Compare:

- **robot minds**
- **minds of software agents**
- **distributed minds**

Placing the study of human minds in an appropriate context

The vast majority of organisms do NOT have human-like architectures, only more or less sophisticated reactive architectures.

(e.g. single-celled organisms, insects, etc.)

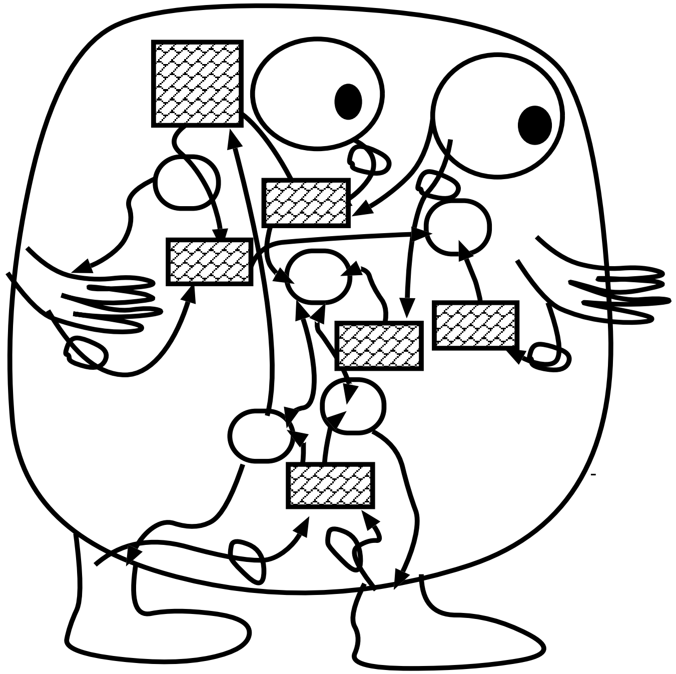
We need to place the study of (normal, adult) human mental architectures in the broader context of:

THE SPACE OF *possible* MINDS

INCLUDING MANY TYPES OF MINDS WITH DIFFERENT ARCHITECTURES THAT MEET DIFFERENT SETS OF REQUIREMENTS, OR FIT DIFFERENT NICHES.

If we don't do this we are likely to attempt to generalise too much from the human case, and get things wrong.

WHAT SORT OF ARCHITECTURE?
Could it be an unintelligible mess?



YES, IN PRINCIPLE.

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

Problem 1:

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.

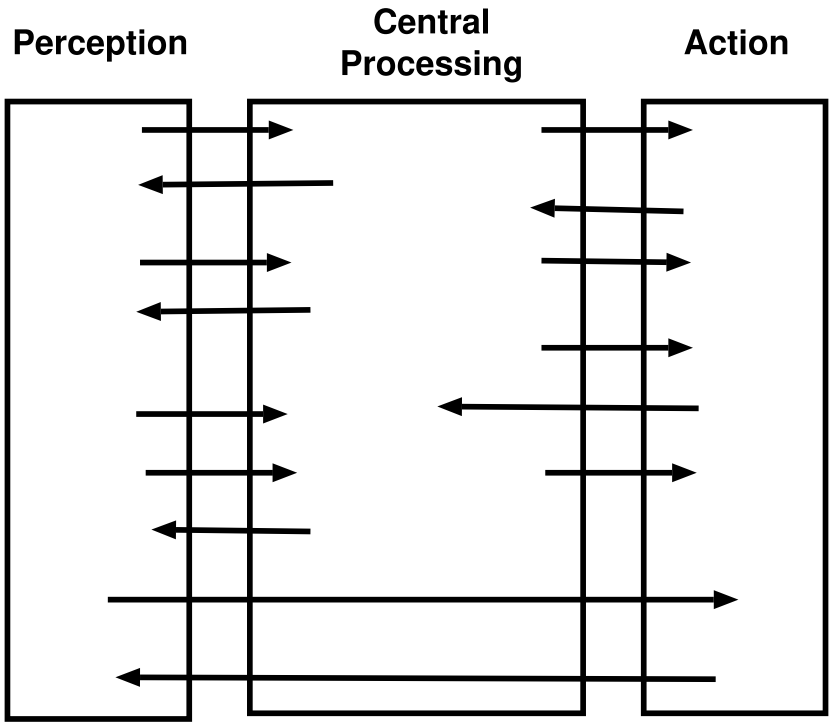
Towards a unifying theory of architectures for natural and artificial agents

1. The "triple tower" perspective

There are many variants,
e.g. Nilsson, Albus....

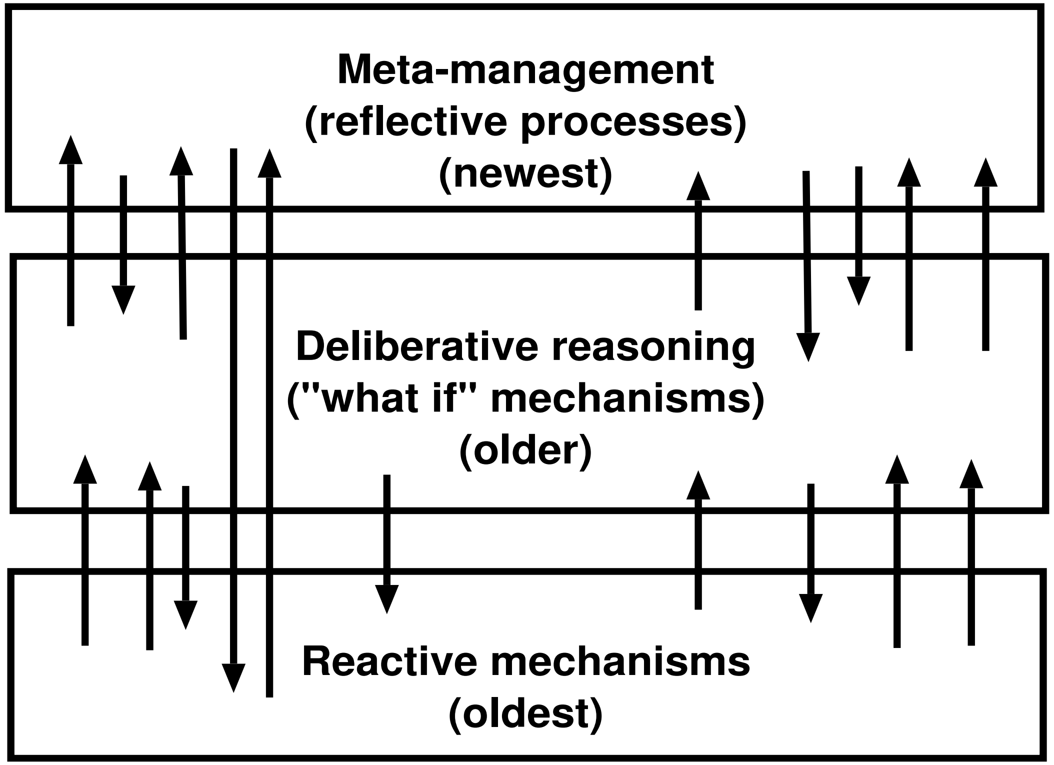
Systems can be
"nearly decomposable".

Boundaries can change with learning and development.



ANOTHER COMMON ARCHITECTURAL PARTITION (functional, evolutionary)

2. The "triple layer" perspective



(MANY VARIANTS – FOR EACH LAYER)

Features of layered systems

- **Reactive** systems can be highly parallel, very fast, and use analog circuits.
- Some reactive capabilities may be innate, others learnt.
- **Reflexes**, with direct connections from sensors to motors, could be separated out from the other reactive mechanisms.
- **Deliberative mechanisms** are inherently slow, serial, knowledge-based, resource limited.
  - Sophisticated deliberative systems require a lot of supporting mechanisms, which may not evolve often, because of their cost.
- **Meta-management** uses additional mechanisms for monitoring, evaluating, and in some cases modifying or controlling internal states and processes.
- In sophisticated organisms meta-management (and other layers) may use culturally determined categories and procedures (e.g. in guilt and self-torment.)

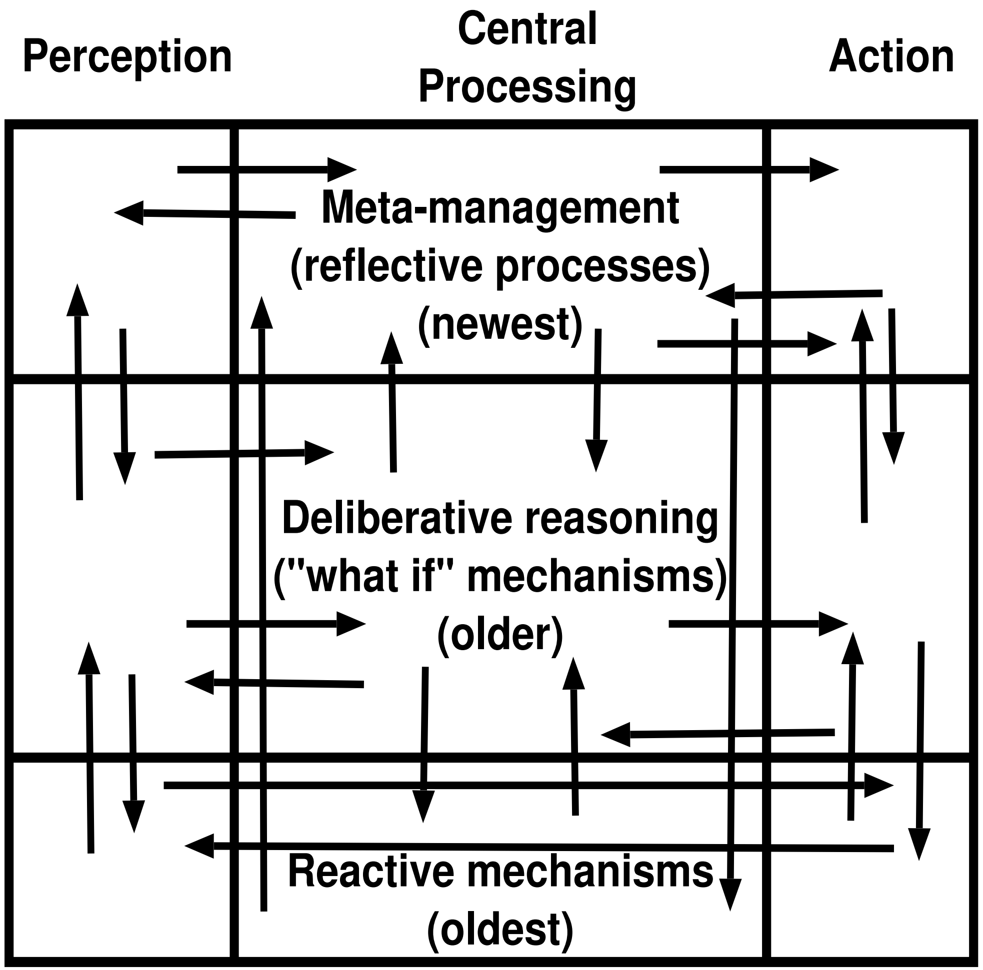
COMBINING THE VIEWS:
LAYERS + PILLARS = GRID

An architectural "schema" (CogAff) not an architecture.

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

Arrows represent some possible routes for flow of information (including control signals).

Other routes are possible, including diagonal routes.



More on the CogAff schema

The CogAff schema defines a variety of types of components, and possible types of information linkages, which may or may not be present in different instances.

To that extent it subsumes a wide variety of possible architectures.

It does NOT specify now control flow must happen, or dominance of control: many options are left open.

Contrast H-Cogaff – a proposed architecture for human-like systems. (Described below).

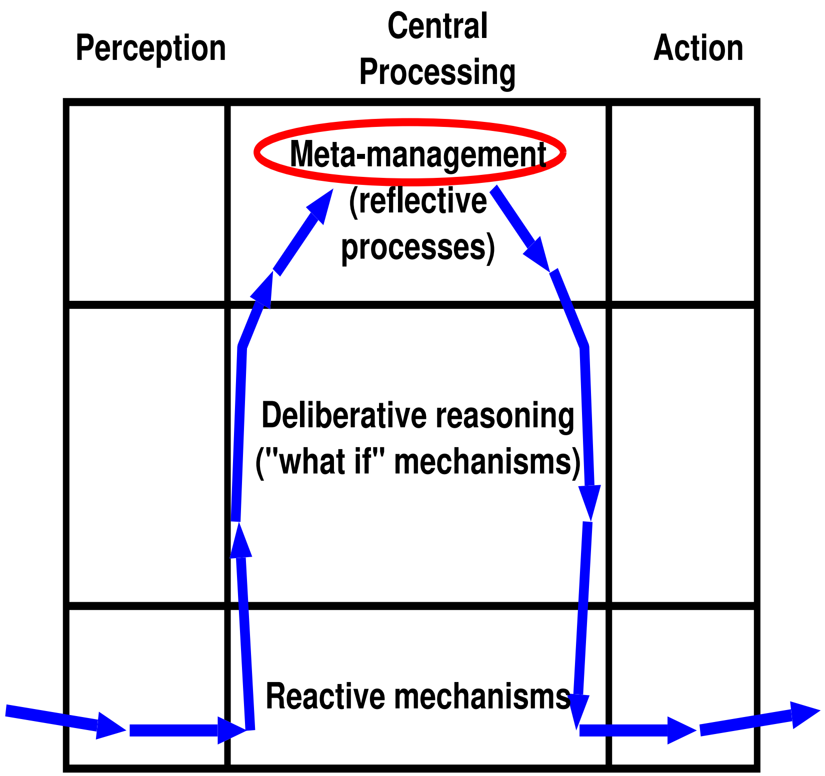
Layered architectures have many variants

With different subdivisions and interpretations of subdivisions, and different patterns of control and information flow.

Different principles of subdivision in layered architectures

- **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, described below.)

The “Omega” model of information flow

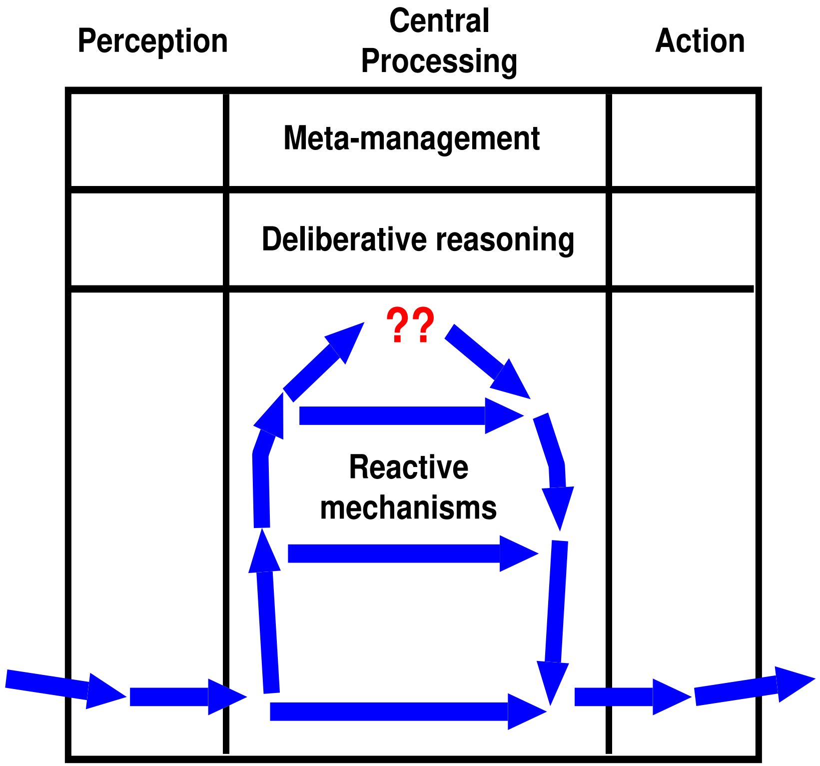


Rejects layered concurrent perceptual and action towers separate from central tower.

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

Some authors propose a “will” at the top of the omega.

Another variant (Brooks): Subsumption architectures



Here all the processing is assumed to be reactive, though there are several layers of reactive processing, including adaptive mechanisms.

Supporters deny that animals (even humans) use deliberative mechanisms.
Yet they somehow get to overseas conferences?

Shakespeare knew that we are information-processing engines:

Love is not love
which alters when it alteration finds

Finding and reacting (or not-reacting) to alteration requires sophisticated information-processing mechanisms.

SENSING AND ACTING CAN BE ARBITRARILY SOPHISTICATED

- Don’t regard sensors and motors as mere transducers.

- They can have sophisticated information-processing architectures.

  E.g. perception and action can each be hierarchically organised with concurrent interacting sub-systems.

  Think of the difference between

  – perceiving edges, optical flow, texture gradients
  – perceiving chairs, tables, support relations
  – perceiving happiness, surprise, anger, which way someone is looking.

Perception goes far beyond segmenting, recognising, describing what is "out there"

It includes:

- providing information about *affordances*
  (Gibson, not Marr, but co-evolved beasties better)
- directly triggering physiological reactions
  e.g. posture control, sexual responses)
- evaluating what is detected,
- triggering new motivations
- triggering "alarm" mechanisms
- . . . . .

AND THESE ALL NEED INTERNAL LANGUAGES OF SOME SORT

An extension of Gibson's theory:

Different sub-systems use different affordances, and different ontologies.

(Evidence from brain damage.)

They rely on processing by different virtual machines: Wittgenstein:

"The substratum of an experience is mastery of a technique"

(In *Philosophical Investigations* Part 2, section xi.

Our work can be viewed as

Steps towards an "ecology of mind"

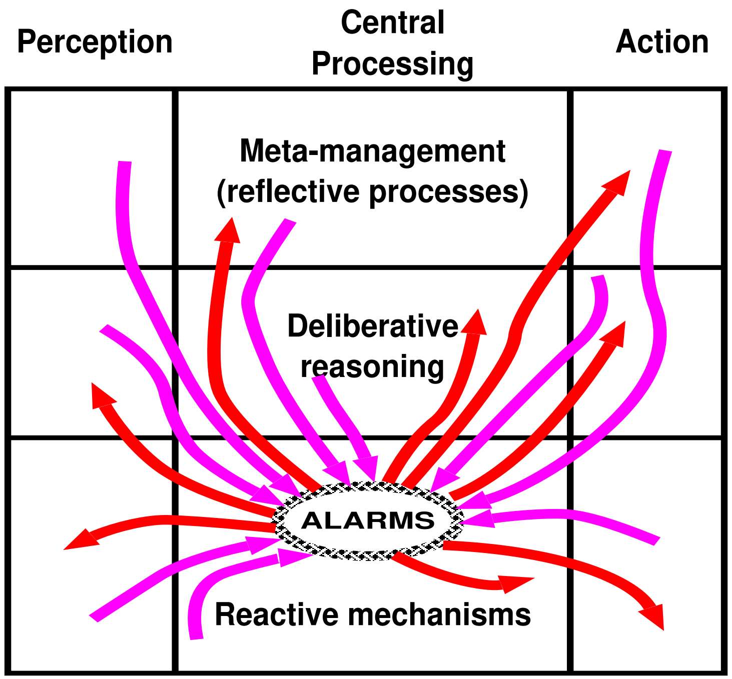
THINK OF A MIND AS AN ECOSYSTEM WITH MANY CO-EVOLVED SUB-SYSTEMS.

As processing grows more sophisticated, so it can be come slower, to the point of danger

REMEDY: FAST, POWERFUL, "GLOBAL ALARM SYSTEMS"

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

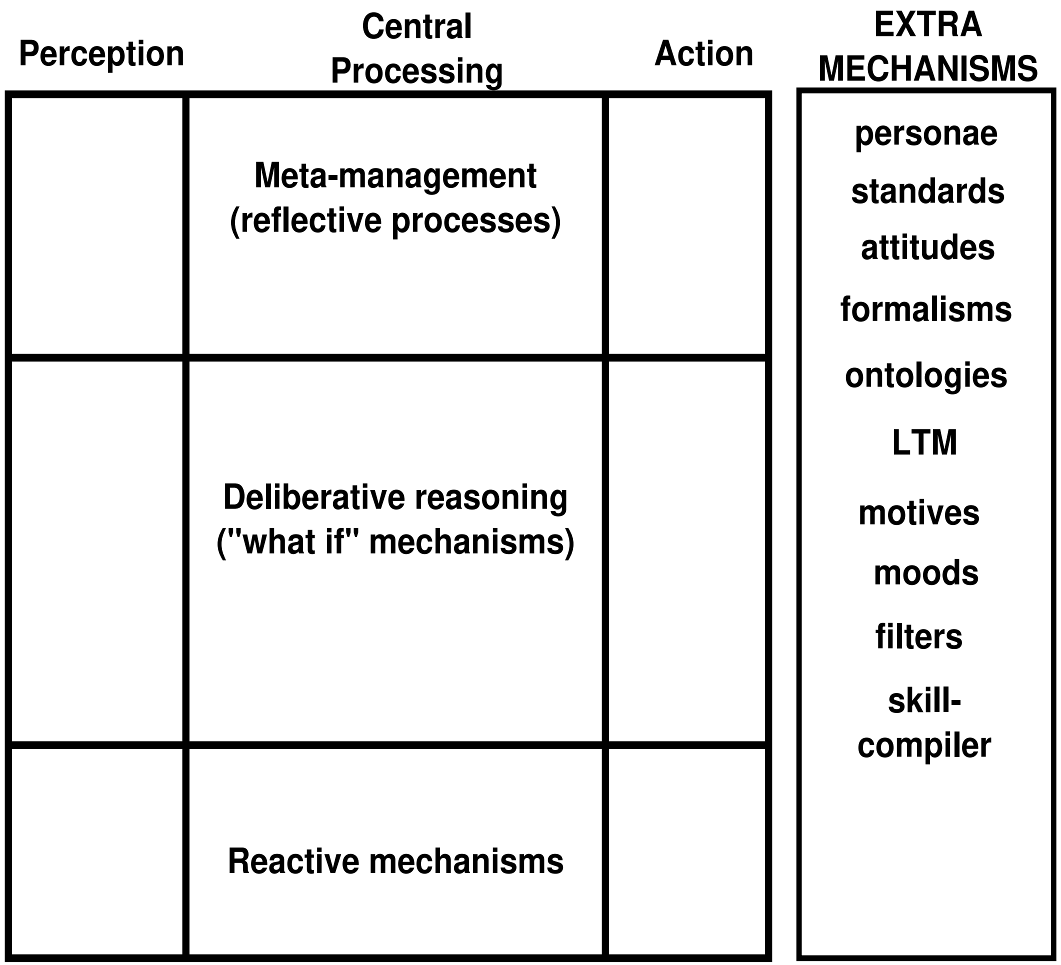
Note: An alarm mechanism is just a reactive subsystem. So it could be seen as part of the reactive layer. Drawing it separately merely serves the pedagogic function of indicating the role.



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

E.g. one alarm system or several? (Brain stem, limbic system, ...???)

ADDITIONAL COMPONENTS



Many profound implications e.g. for kinds of development, kinds of perceptual processes kinds of brain damage, kinds of emotions. (No time to discuss fully)

VARIETIES OF MOTIVATIONAL SUB-MECHANISMS

MOTIVATION IS NOT JUST ONE THING

- Motives or goals can be short term, long term, permanent.
- They can be triggered by physiology, by percepts, by deliberative processes, by metamanagement.
- They can be implicit in the operation of active mechanisms, or explicit.
- They can be part of the reactive system, part of the deliberative system, part of meta-management.

Motive generators

There are many sorts of motive generators: MG

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

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

- Motive generator generators: MGG
- Motive comparator generators: MCG
- Motive generator comparators: MGC

  and maybe more:

    MGGG, MGGC, MCGG, MCGC, MGCG, MGCC, etc ?

There are also EVALUATORS

Current state can be evaluated as good, or bad, to be preserved or terminated.

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

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

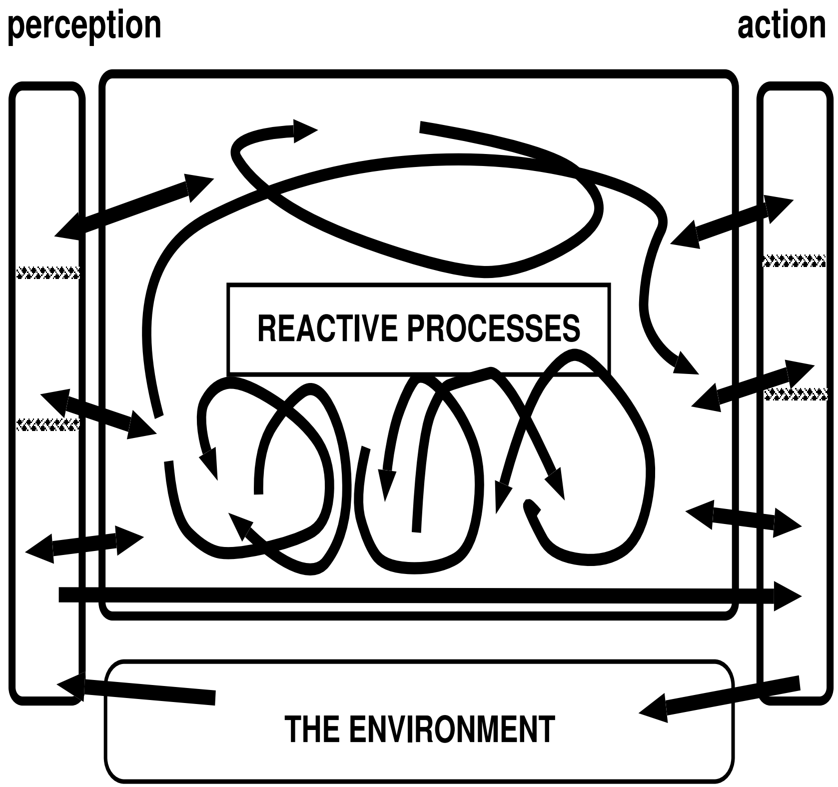
Evaluations are often confused with emotions.

But something can be evaluated as good or bad quite unemotionally (coldly).

A special case of evaluation: “error signals” e.g. during feedback control.

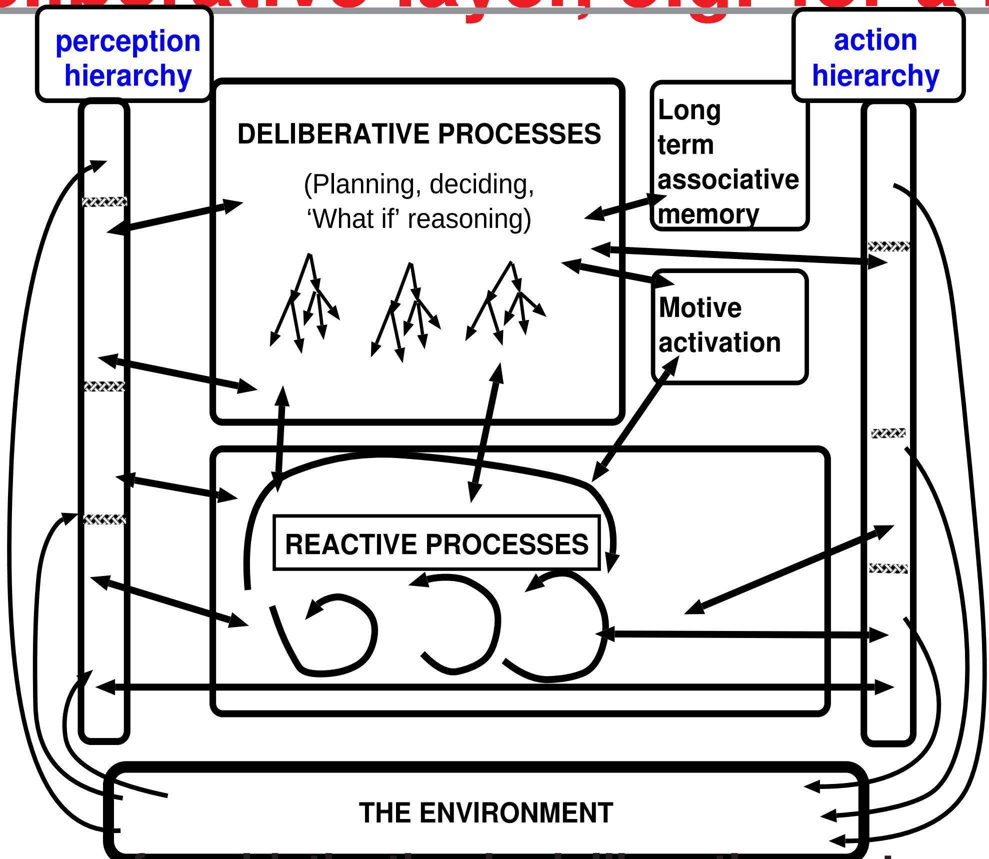
NOT ALL PARTS OF THE GRID ARE PRESENT IN ALL ANIMALS

How to design an insect?



Purely reactive but can include adaptation and internal state monitoring.

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

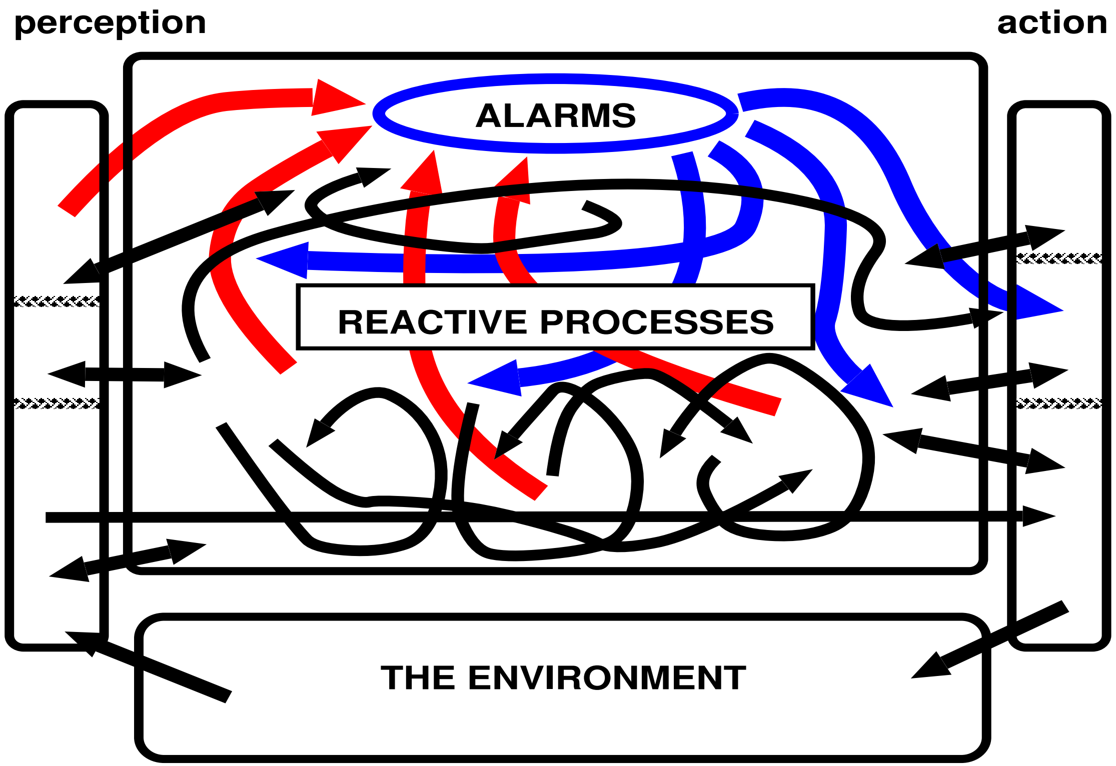


There various degrees of sophistication in deliberative systems.

The key requirement for the most sophisticated version is a formalism and mechanisms for doing hypothetical reasoning.

The formalisms and mechanisms in other animals are probably much more restrictive than in humans.

EMOTIVE INSECTS?
(with alarm mechanisms)



Alarm mechanisms allow rapid global redirection of processing when certain patterns are detected.

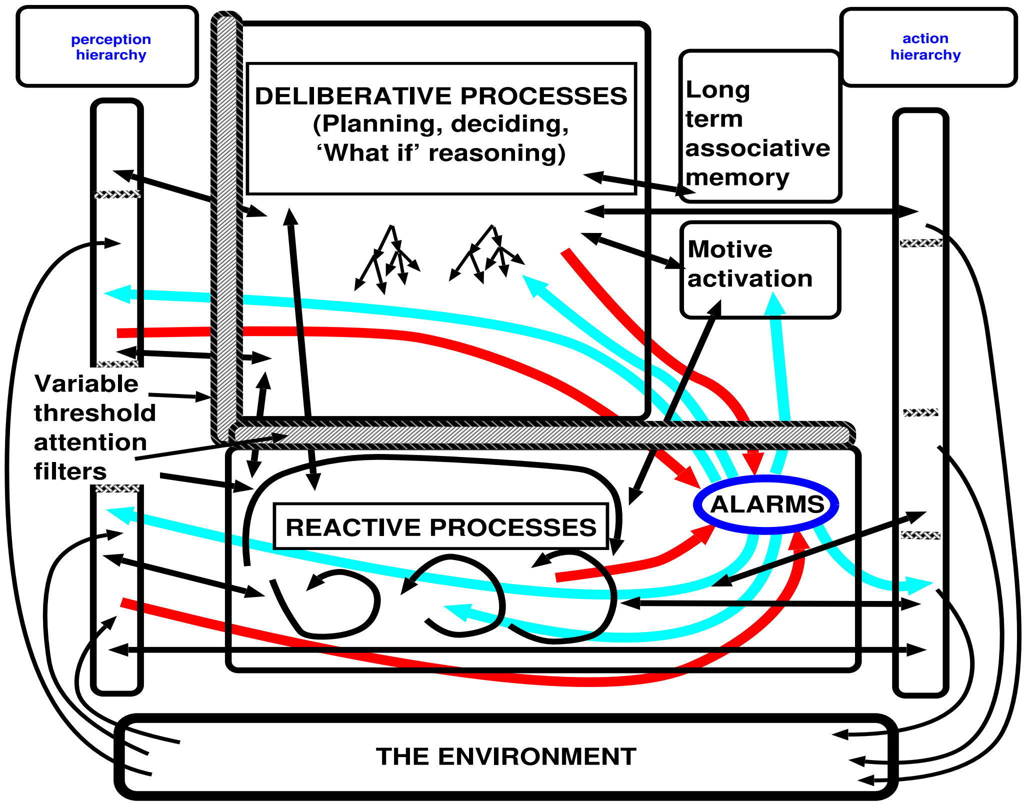
Different versions support different primitive emotions.

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.

Secondary emotions

What Damasio and Picard call "Secondary Emotions" seem to be reactions triggered by central cognitive processes in a deliberative mechanism.

Note: Whether these involve the same physiological responses as primary emotions in humans and other animals is an empirical question.

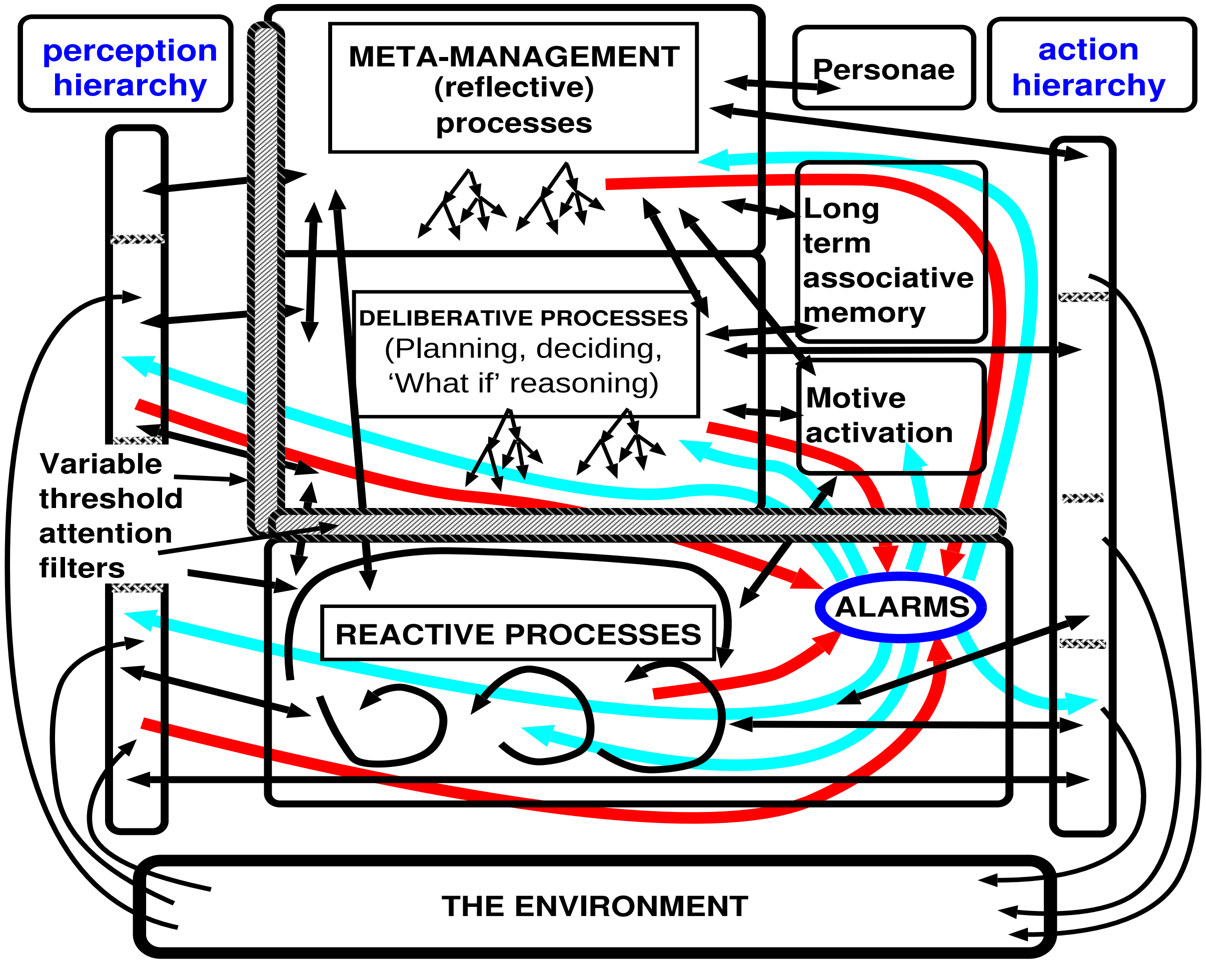
There is no *theoretical* reason why they should *always* do so.

Humans seem to vary in this respect.

E.g. training can suppress normal reactions even though the emotion persists as an internal state (often highly dispositional).

It is important to distinguish the *empirical* question whether in most or all humans certain central processes produce certain bodily changes from the *conceptual* question whether the occurrence of those bodily changes has to be taken as part of the DEFINITION of a type of state, e.g. an emotion. That would rule out apriori aliens whose grief, jealousy or joy lacked our physical side-effects. There does not appear to be any good scientific basis for such dogmatism.

H-COGAFF: A human-like architecture.



H-COGAFF is a special case of the CogAff schema using all the components

Described in more detail in papers in the Cogaff directory:

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

ONE OR MORE ALARM MECHANISMS
(Brain stem, limbic system, blinking reflexes, ...???)

Alarm mechanisms in H-CogAff allow rapid redirection of the whole system or specific parts of the system required for a particular task (e.g. blinking to protect eyes.)

They 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.

Tertiary emotions

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

Involve interruption and diversion of thought processes. I.e. the metamanagement 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.

In particular, tertiary emotions in which control of attention is disrupted do not require physiological changes in the body.

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 – mistakenly accepted by many researchers. Perhaps wishful thinking: it would be "nice" to think that emotions are needed for intelligence?)

Different architectural layers support different sorts of emotions, and help us define architecture-based ontologies for different sorts of minds

Different animals will have different mental ontologies

Humans at different stages of development will have different mental ontologies

THE THIRD LAYER

enables

SELF-MONITORING, SELF-EVALUATION
and
SELF-CONTROL

AND THEREFORE ALSO LOSS OF CONTROL (TERTIARY EMOTIONS: PERTURBANCES)

and qualia (through concurrent self-monitoring e.g. of sensory databases)!

NOTES:

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

2. Remember that these are not static states but developing processes, with very varied aetiology.

Different forms of development correspond to different sorts of emotions.

3. We don't necessarily already have names for all the significantly different cases

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

SOCIALLY IMPORTANT HUMAN EMOTIONS

INVOLVE RICH CONCEPTS AND KNOWLEDGE AND HIGH LEVEL CONTROL MECHANISMS (architectures)

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.

SUMMARY

1. We can reduce conceptual muddles regarding emotion, etc. by trying to use architecture-based concepts.

2. Different architectures are relevant in different contexts (e.g. infants, adults, other animals). So we need to explore different families of concepts (e.g. for describing infants, chimps, cats, people with brain damage).

3. Finding out which architectures are relevant is a hard research problem. We suggest that humans have three architectural layers that manifest themselves not only centrally but also in perception and action sub-systems. Most other animals have only a subset.

4. At least three (and several more if we look closely) classes of affective states and processes can be distinguished, related to different architectural layers.

5. Many other concepts (e.g. “learning”, “belief”, “motivation”, “intentional action”) can be refined on the basis of hypothesised architectures.

Summary continued

6. The complexity and variety of affective states and processes supported by the H-Cogaff architecture can explain some of the confusion in the literature: different researchers focus on different subsets. hence their definitions and theories are different.

8. The ability to have emotions is a *side-effect* of mechanisms required for intelligence (as argued in IJCAI-1981). Contrast the illogical use of facts about frontal lobe damage to infer that intelligence *requires* emotions.

CONCLUSION

- Much of this is conjectural – many details still have to be filled in and consequences developed (both of which can come partly from building working models, partly from multi-disciplinary empirical investigations).
- An architecture-based ontology can bring some order into the morass of studies of affect (e.g. myriad definitions of “emotion” are explained as based on partial views).
- This can lead to a better approach to comparative psychology, developmental psychology (the architecture develops after birth), and the study of effects of brain damage and disease.
- The CogAff schema provides a conceptual framework for discussing which kinds of emotions can arise in various kinds of artificial agents, e.g. software agents that lack the reactive mechanisms required for controlling a physical body.
- All this may be relevant not only to science, but also to ambitious engineering objectives listed at the beginning.

THE BIRMINGHAM COGNITION AND AFFECT PROJECT

OVERVIEW:

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

PAPERS:

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

http://www.cs.bham.ac.uk/~axs/cogaff/simagent.html
(the SIM_AGENT toolkit)

SLIDES FOR TALKS:

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

(including this talk)