Presented at Symposium on Architectures – Stanford, March 2003,
Presented at University of Notre Dame April 2003

Varieties of affect and learning in a complete human-like architecture

AARON SLOMAN

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

School of Computer Science
The University of Birmingham

This slide presentation can be found at
http://www.cs.bham.ac.uk/˜axs/misc/talks/#talk24

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

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

THANKS

- To Matthias Scheutz and colleagues and students in the Cognition and Affect project at the university of Birmingham.
  http://www.cs.bham.ac.uk/ axs/cogaff.html

- To Marvin Minsky for inspiration since I discovered AI as the best way to do philosophy in 1969, e.g. his 'Steps towards Artificial Intellingence' (1961) and *The Emotion Machine*, draft book online at his web site.

- To the Leverhulme Trust for research support

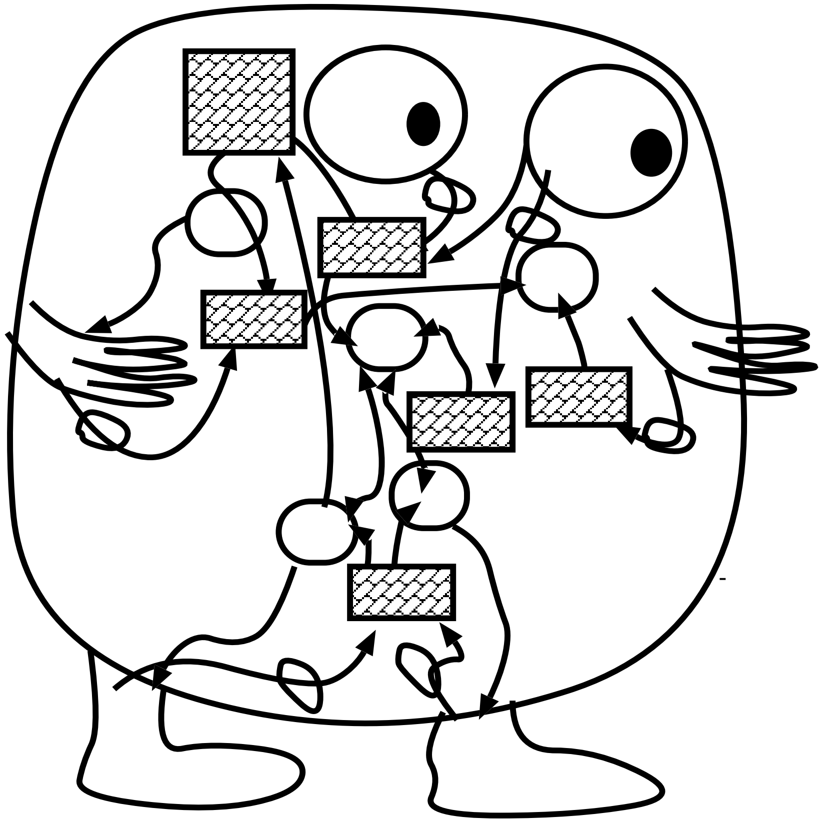
To the developers of Linux
and other free, portable, reliable, software systems,
e.g. Latex, Tgif, xdvi, ghostscript, Poplog, etc.

No Microsoft software required.

WHAT IS AN ARCHITECTURE?
The good side of "boxology"

- A specification for an integrated functioning system, in terms of the components within it (which may have their own architectures) and the causal and communicative links between those components which explain how the whole system works.
- It may include different forms of representation in different parts of the system.
- The components may be either
  - physical machines

    or
  - virtual machines.



There may also be

- sensory transducers
- motor transducers (effectors)

Examples of virtual machine architectures

Running versions of the following are virtual machines, within which events and processes occur, objects are created, modified, destroyed:

- Operating systems
- Word processors
- ACT-R
- Soar
- The internet
- Social systems
- ecosystems

There are many architecture-wars in AI and Cognitive Science

For more on the nature of virtual machines and their causal powers see

http://www.cs.bham.ac.uk/research/cogaff/talks/#talk23
http://www.cs.bham.ac.uk/~axs/ijcai01/

Abstract

I’ll introduce

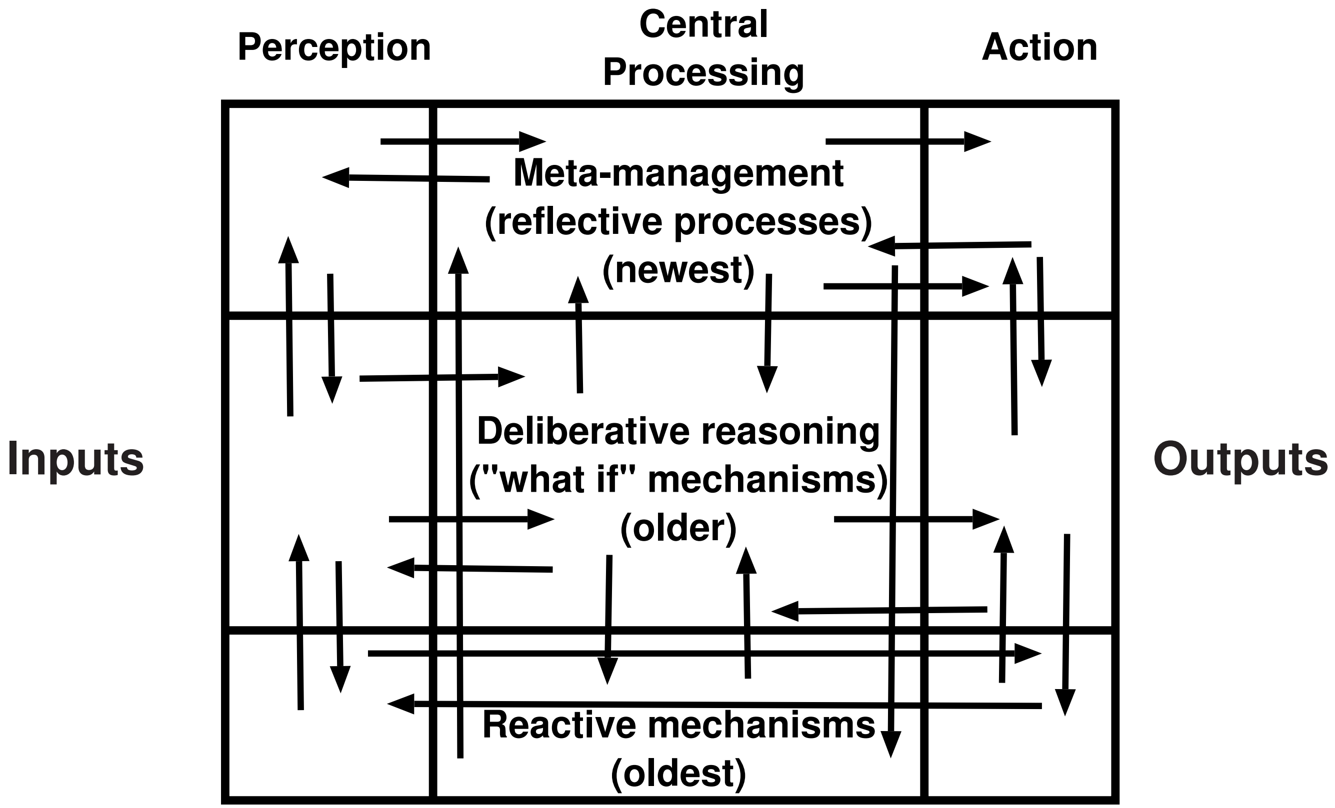
- CogAff, a framework for thinking about integrated architectures combining different forms of representation and different mechanisms.

  This may provide a useful framework for comparing and evaluating different architectures

- H-Cogaff, a special instance of CogAff — a conjectured architecture for human-like systems, incorporating
  - diverse concurrently active components
  - for an integrated agent
  - perceiving and acting
  - in a complex, changing, partially understood world,
  - including other intelligent agents.

The CogAff Framework (Schema) Summarised

A sort of generative grammar for a class of architectures for integrated agents, embedded in a complex and changing environment, which is perceived and acted on:

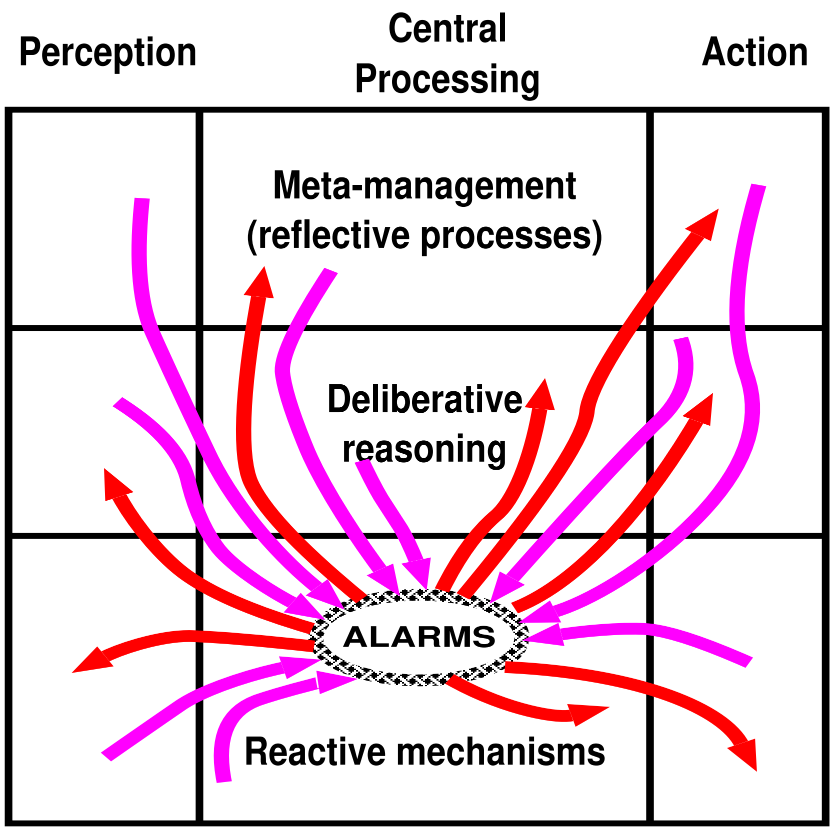


Not all possible information flows are shown!

Different architectures include mechanisms in different subsets of the boxes, and different possible information links, different possible control relationships.

There are also differences in forms of representation and types of mechanisms.

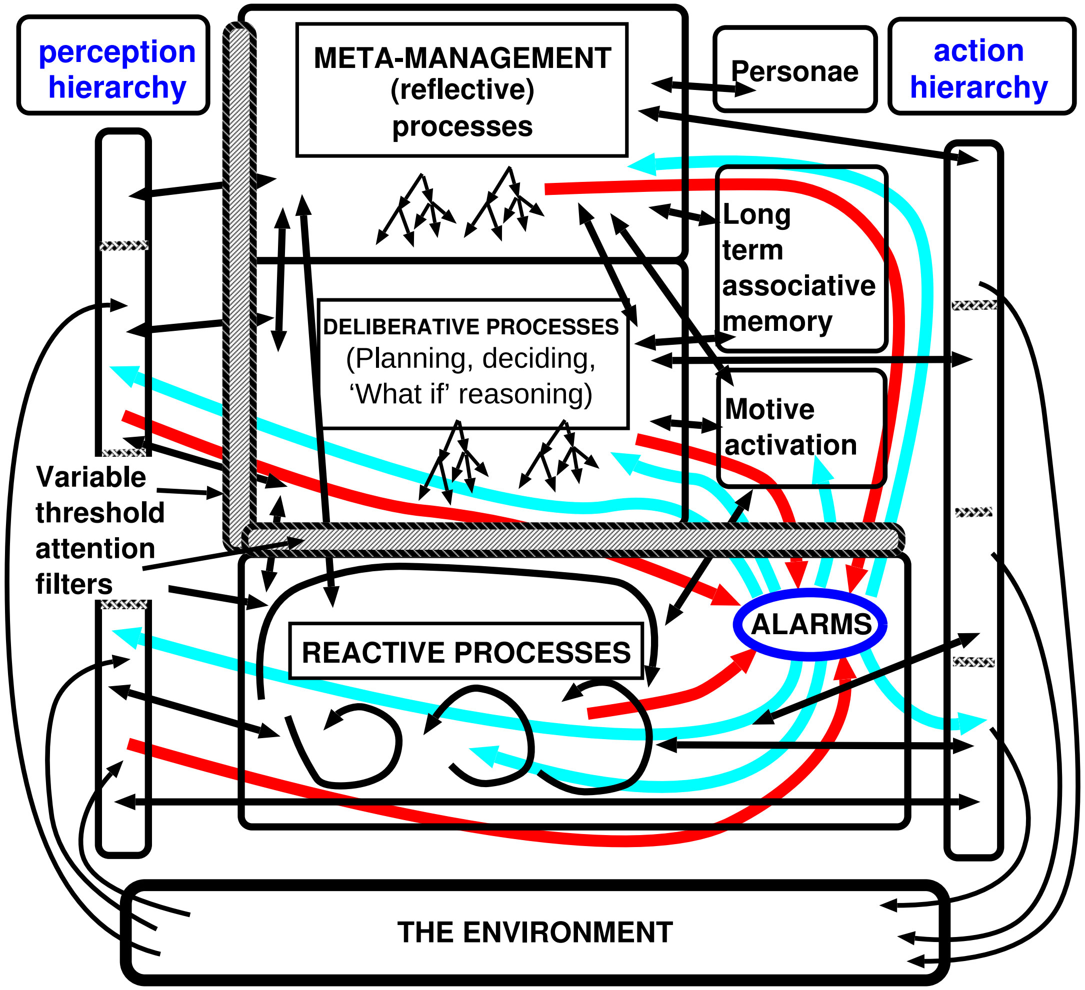
With "alarm" mechanisms for rapid global redirection



The architectural basis for a subset of emotions

Sketch of the H-Cogaff Architecture

One of many architectures compatible with the CogAff Schema, explained later.



NB: emergent emotional states, but no 'emotion box'.

Conjectures: Part 1

- Something like the CogAff schema can enable our field to make better progress by providing a common ontology and vocabulary for comparing and contrasting architectures.

- Something like the H-Cogaff architecture can accommodate many features of human mental functioning, including emotions and learning.

  It allows for

  – more varieties of emotions and more varieties of learning than are normally considered

  – diverse affective states, including desires, pleasures, pains, attitudes, moods, etc.

  – maybe even “qualia” solving old philosophical problems about consciousness.

But I don’t have time to explain all this today.

Conjectures: Part 2

- **Architecture-based concepts** defined in terms of H-Cogaff, can refine and extend many of our *very confused* concepts,
  e.g. emotion, learning, consciousness (cluster concepts).
- Related Architecture-based concepts, using other instances of the CogAff framework, will be applicable to other organisms and machines,
  e.g. "emotions" in insects.
- Models based on H-Cogaff can play a useful role both for
  - applications of AI (e.g. digital entertainments, or in learning environments), and for
  - scientific theories about human minds and brains.

NOTE:

At present the specifications are extremely sketchy.
Some details could be filled in using known AI techniques and formalisms.
Most of it requires major future advances, some discussed in other slide presentations.

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

Vision, especially, is largely not understood: e.g. we don't know what its functions are. I give some partial answers later.

METHODOLOGICAL POINT 1

We cannot expect deep understanding if we merely develop one architecture without exploring:

- **Design space: the space of possible architectures**
  (or at least a relevant neighbourhood in that space)

  Compare John Barrow's book *Constants of Nature*
  (exploring alternative versions of general physical/cosmological theories).

- **Niche space: the space of possible sets of requirements for architectures**
  (or at least a relevant neighbourhood in that space)

- **Relationships between (the relevant neighbourhoods in) design space and niche space.**

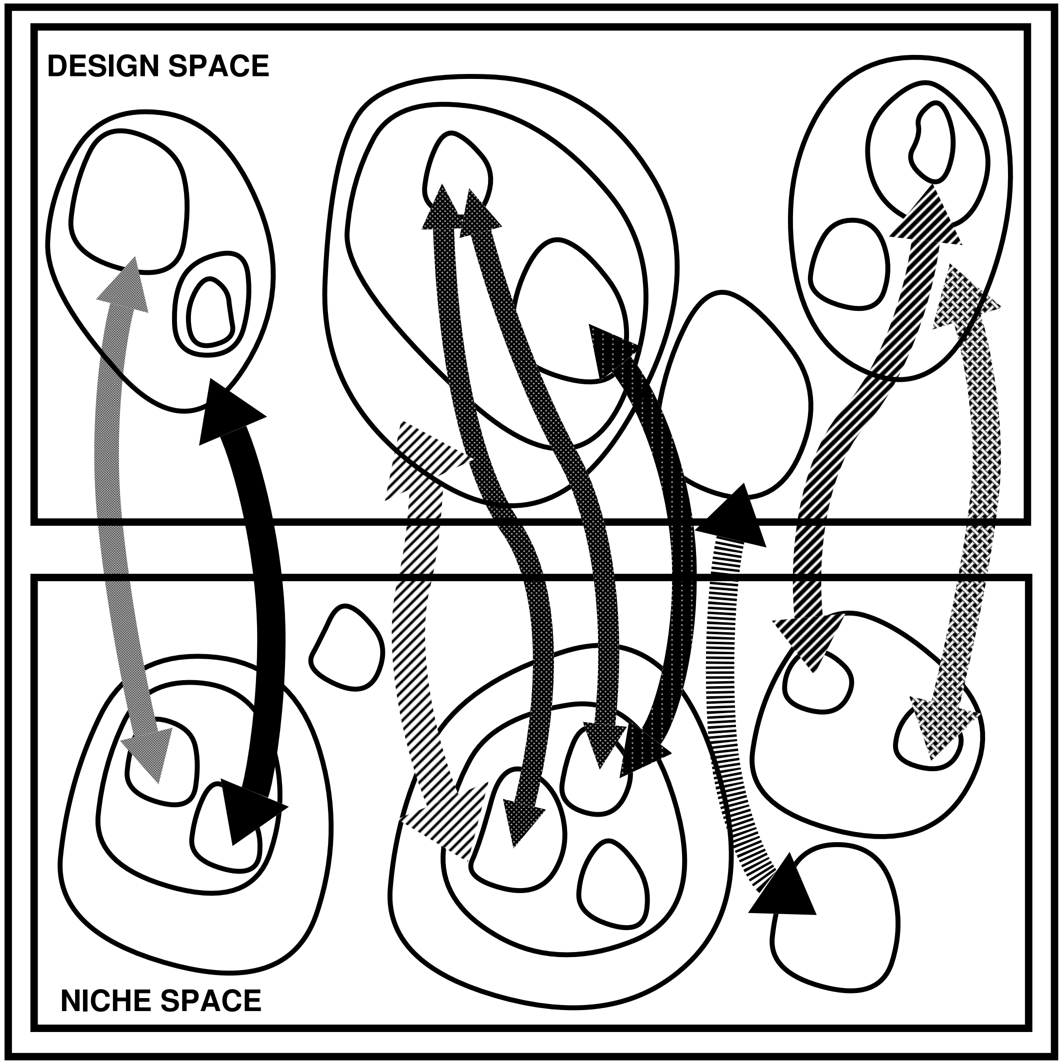
However we can sometimes solve practical problems without deep science. (See http://www.cs.bham.ac.uk/research/cogaff/talks/#talk4)

Design space and niche space

Design space: the space of possible architectures.

Niche space: the space of possible sets of requirements for architectures.

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



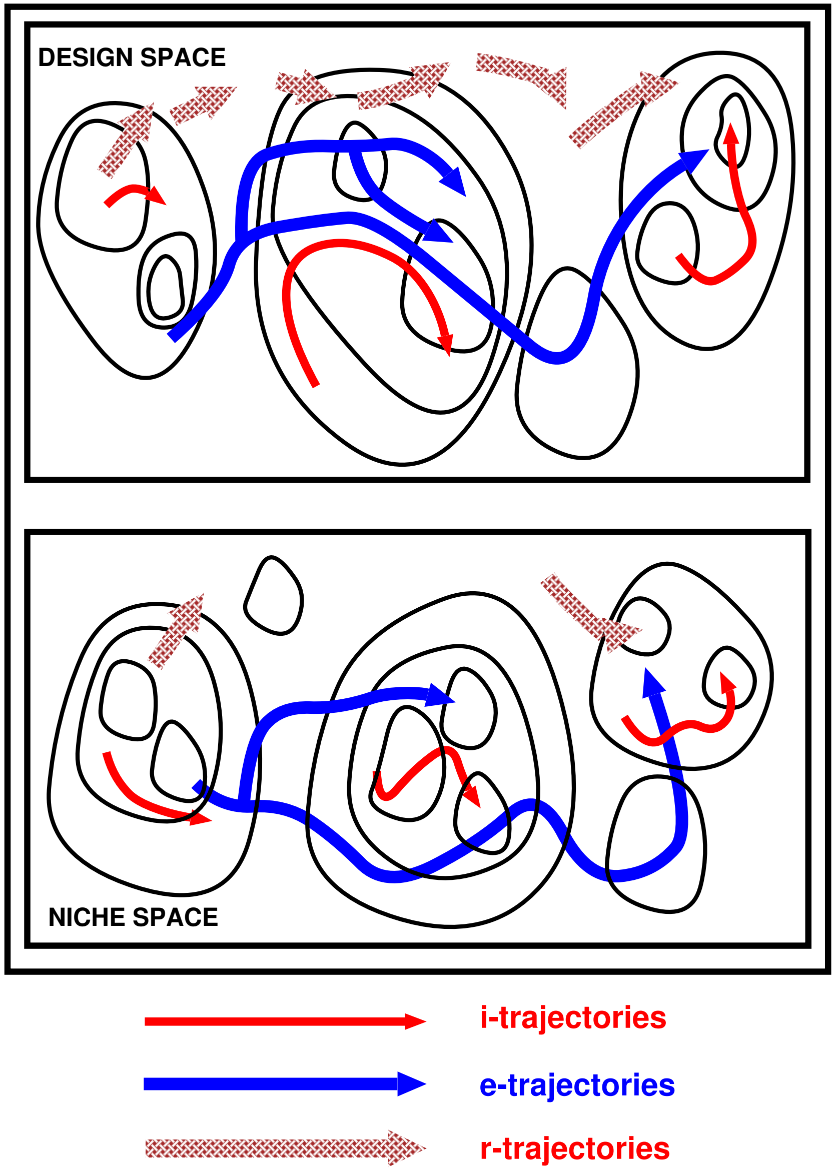
Do not expect one fitness function.
Instead expect *diverse structured fitness relations* between designs and niches.

Trajectories in design space and niche space

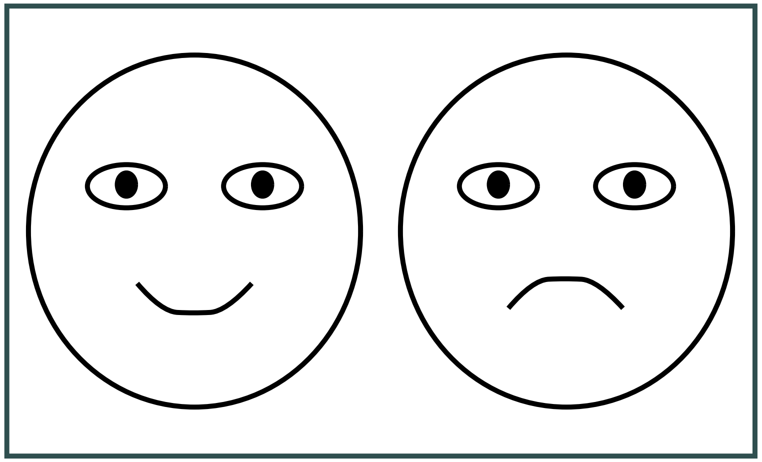
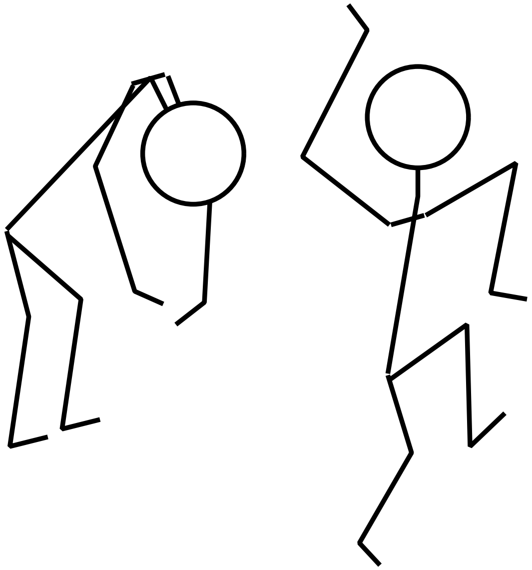
There are different sorts of trajectories in both spaces:

- **i-trajectories:**
  Individual learning and development
- **e-trajectories:**
  Evolutionary 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 (not shown separately).



What do we mean by having an emotion?

- Is it enough to produce certain behaviours that people interpret as emotional?
- Does it matter what the architecture underlying the behaviour is?

- Behaviour is not enough to define any mental state, since
- in principle any behaviour can be produced by indefinitely many different mechanisms,
- including some in which all behaviour is fully scripted in advance by the designer,
- unlike thought and behaviour in humans and many other animals.

  e.g. Betty the crow (perhaps): http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm

METHODOLOGICAL POINT 2

The concept of emotion is but one of a large family of intricately related concepts, including many affective concepts.

E.g. moods, attitudes, desires, dislikes, preferences, values, standards, ideals, intentions, etc., the more enduring of which (along with various skills and knowledge) can be thought of as making up a "personality".

Models that purport to account for emotion without accounting for others in the family are bound to be shallow

though they may have practical applications.

(See http://www.cs.bham.ac.uk/research/cogaff/talks/#talk3)

A "periodic table" for affective concepts can be based on an architecture, in something like the way the periodic table of elements was based on an architecture for physical matter.

The analogy is not exact: there are many architectures, each providing its own family of concepts.

There may be some concepts applicable across architectures

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.

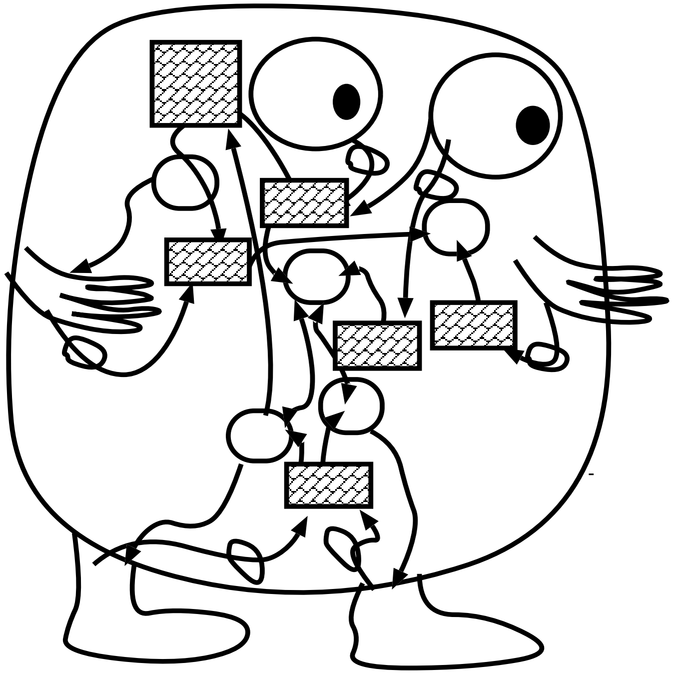
NOTE: I am not specially interested in *emotions* except as a special case of a wide range of phenomena that need to be accommodated in a theory of what minds are (an ontology for minds) and explanations of how they work.

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)

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

YES, IN PRINCIPLE.
As some have argued.



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.

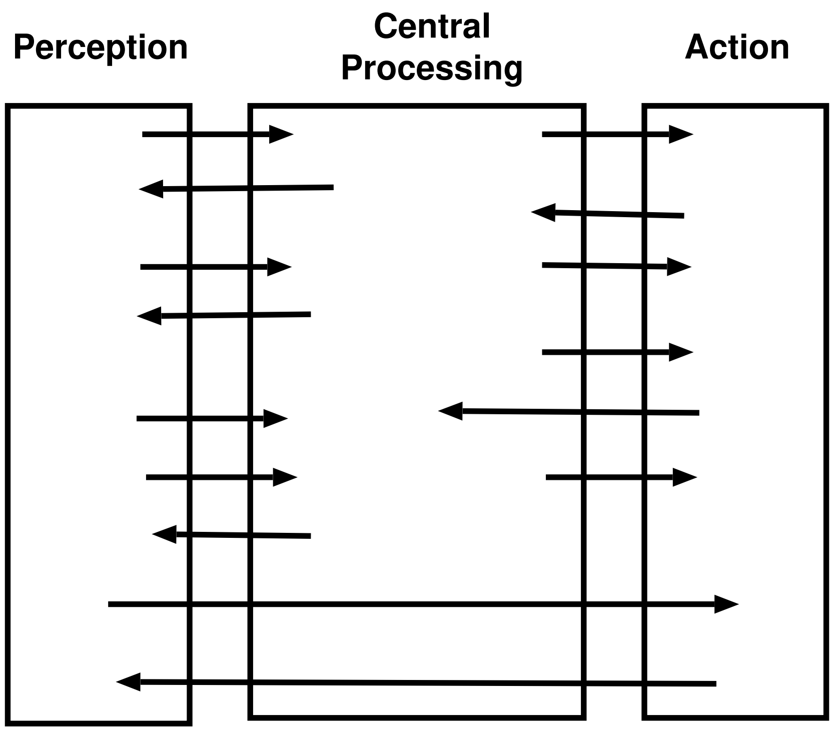
Towards a unifying theory of architectures for natural and artificial agents

1. A "triple tower" perspective

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

Systems can be
"nearly decomposable".
(Herbert Simon)

Boundaries can change with learning and development.

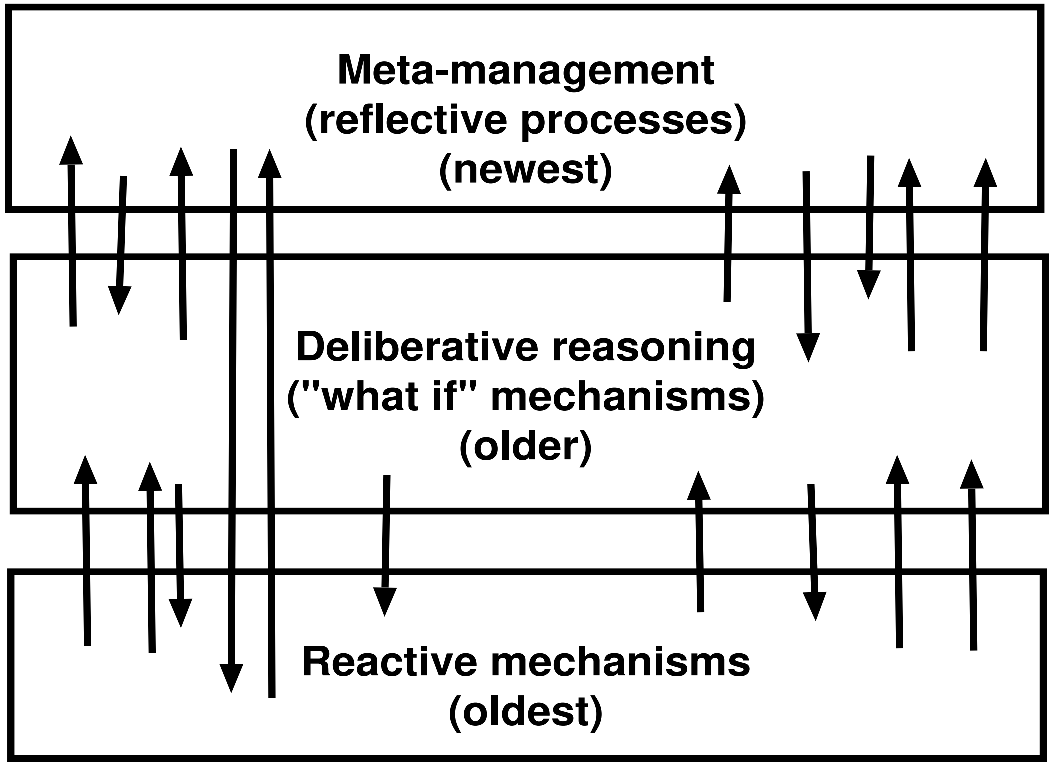


Many theories fail to do justice to the complexities of perception and action, because people do not analyse the requirements in depth.

There can be *hierarchies* of percepts and actions.
See http://www.cs.bham.ac.uk/research/cogaff/talks/#talk21

ANOTHER COMMON ARCHITECTURAL PARTITION (functional, evolutionary)

2. A “triple layer” perspective



(MANY VARIANTS – FOR EACH LAYER)

This is not the “three layer”system of Eran Gatt.

Features of the layers

- **Reactive** systems can be highly parallel, very fast, and use a mixture of analog circuits and digital, e.g. rules.
  - 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, with powerful formalisms for expressing descriptions of alternative possibilities, require a lot of supporting mechanisms, which may not evolve often, because of their cost, e.g. requiring "expensive" brain mechanisms (at peak of food pyramid)
- **Meta-management** uses additional mechanisms for monitoring, categorising, 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.)

The layers may be concurrent or interleaved or pipelined.
There may or may not be a dominance hierarchy.

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,**
- **forms of representation used**
- **kinds of mechanisms used**
- **which bits use analog mechanisms, which digital**
- **how much concurrency**
- **whether components are synchronised**
- **how many concurrent goals can exist**
- **how goal conflicts are resolved (arbitration, preferences....)**
- **control-hierarchy,**
  (Top-down vs multi-directional control. See subsumption architectures, below.)
- **information flow**
  (e.g. the popular 'Omega' Ω model of information flow, described below.)

Beware of terminological differences

E.g.

- Some people use “reactive” to mean stateless – we don’t.
- Some people use “reactive” to exclude chained automatic routines – we don’t.
- Some people use “reactive” to exclude the use of discrete symbols – we don’t.
- Some people use “reflective” to refer to “watchful” processes of plan-execution leading to plan improvements, etc. – we include that under “deliberative.”
- Our top layer (“meta-management”) requires special architectural support for an internal process observing, categorising, evaluating, and possibly controlling or modulating other internal processes.
  Perhaps it could evolve by copying and redeploying “alarm” mechanisms.

COMBINING THE VIEWS: LAYERS + PILLARS = GRID

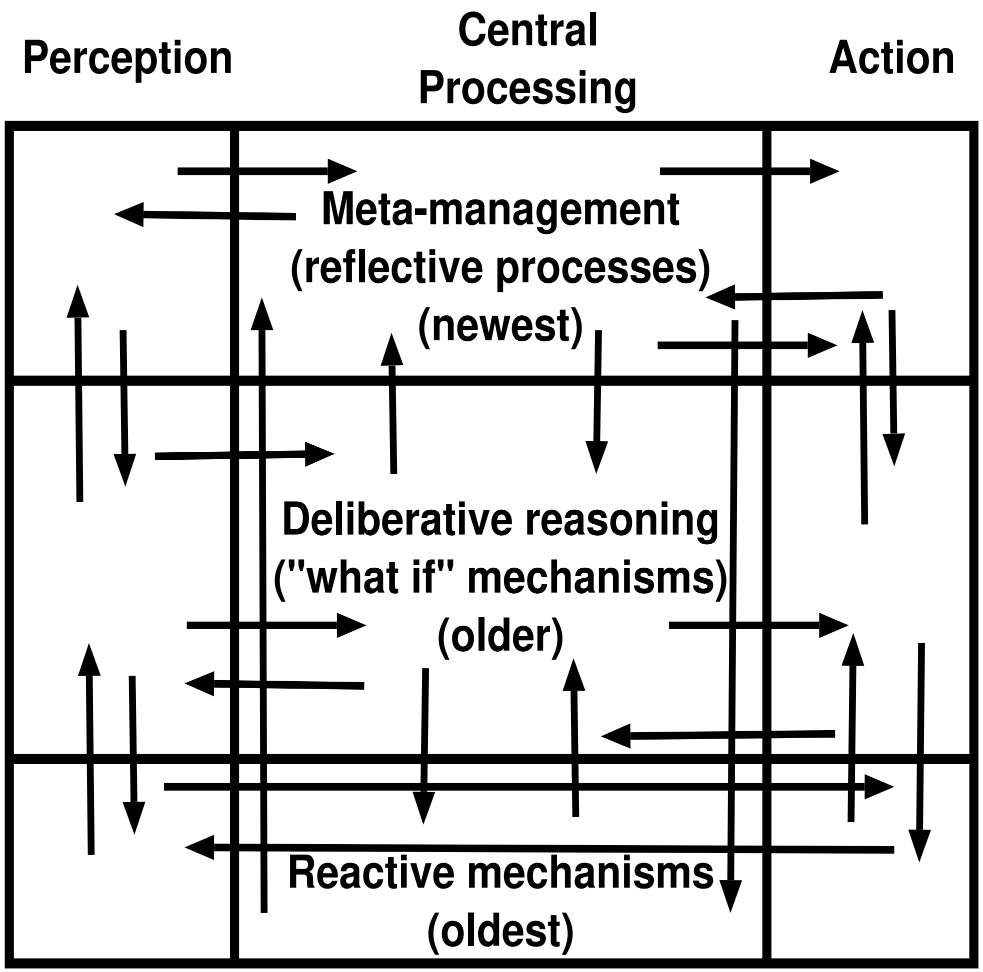
An architectural "schema" (CogAff) not an architecture.

A grid of co-evolved 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.

Not all organisms will have all the kinds of components, or connections.



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.

  We understand very little about what affordances are, how they are represented, how they are perceived, how they are used.

  A draft paper is here:

  http://www.cs.bham.ac.uk/research/cogaff/sloman-diag03.pdf

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

It includes:

- providing information about *affordances*
  (Gibson was closer to the truth than Marr)
- 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

Multi-window perception

We'll propose a "multi-window" theory of vision (and action, and possibly other modalities of perception.)

- The key idea is that visual mechanisms operate on different levels of abstraction in parallel, combining bottom up, top down, and background information in very flexible ways.
- The different levels may use different ontologies to characterise what is perceived.
- Some of them can be concerned with evaluation as well as description.
  (Compare L.Pryor's "reference features")
- Contrast "peephole" perception: sensory information is a homegeneous collection of information processed in homogeneous ways (e.g. statistical methods).

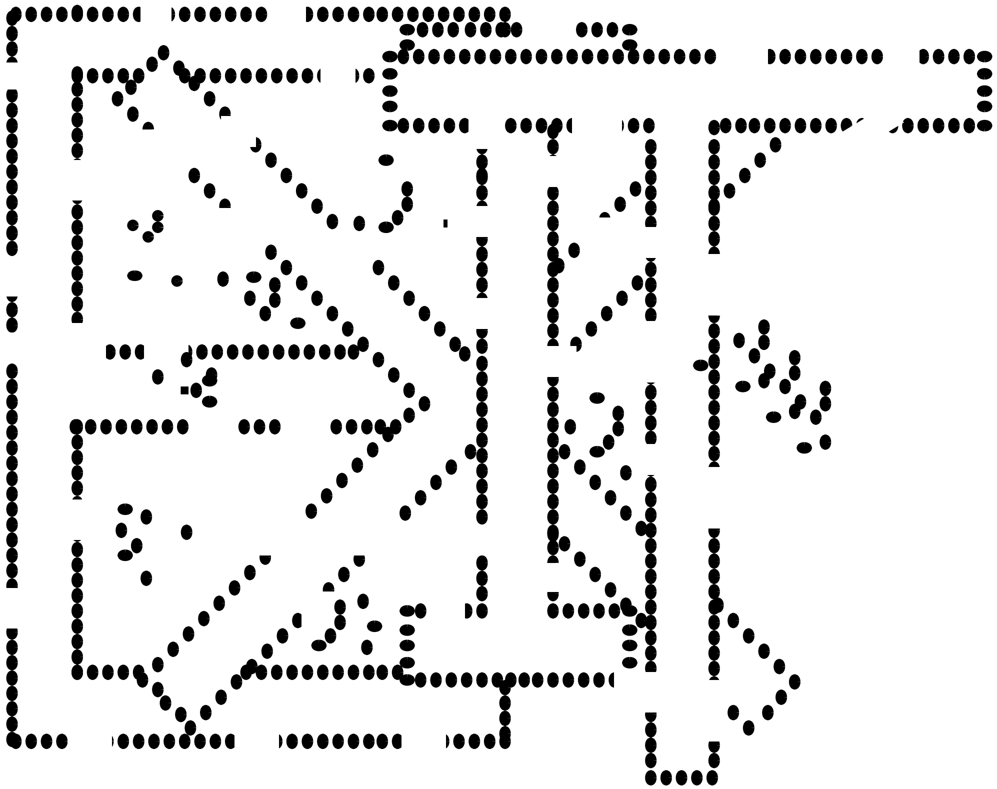
Architectural theories which ignore the possibilities of multi-window perception and action will fail to account for some of the complexities of human minds.

This may not matter for some engineering applications.

What do you see?

Despite all the clutter, most people see something familiar.

Some people recognize the whole before they see the parts.



Animal visual systems are not presented with neatly separated images of individual objects, but with cluttered scenes, containing complex objects of many sorts often with some obscuring others. The objects may be moving, may be hard to see because of poor lighting, or fog, or viewed through shrubs, falling snow, etc.

Real seeing is often much harder than the tasks most artificial vision systems can perform at present.

It is also, in its way, much harder than many of the tasks presented by psychologists to subjects in vision laboratories

(selected for suitability for repeatable laboratory experiments)

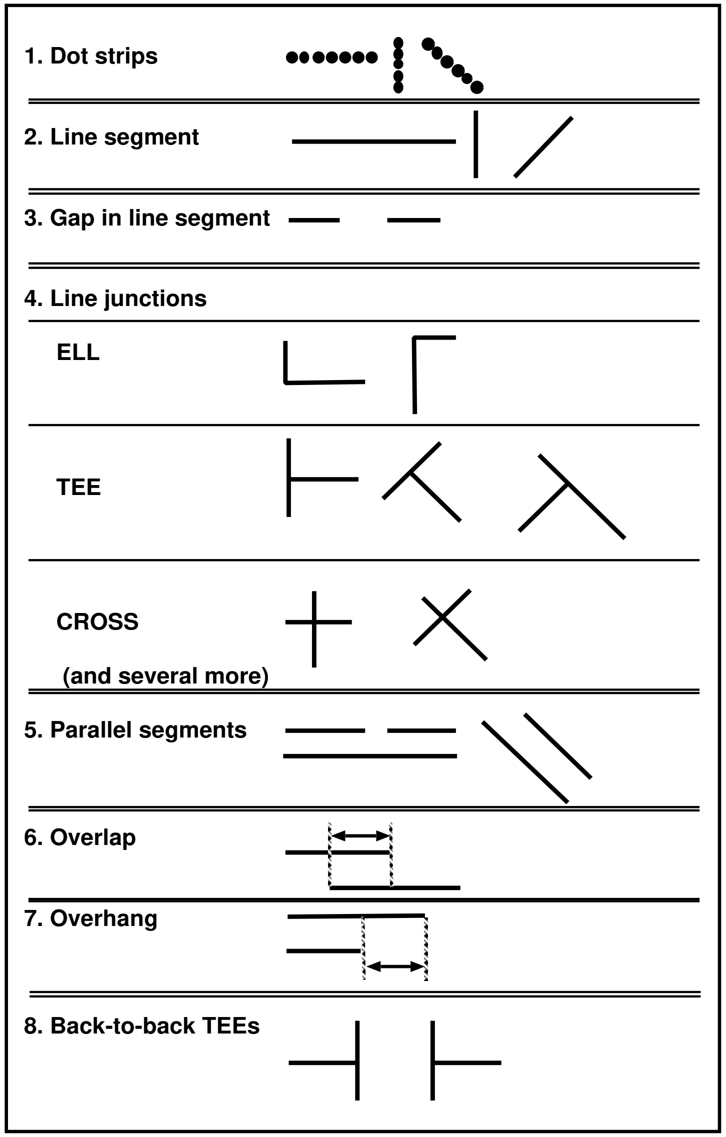
Conjecture:
Assemble fragments at different levels of abstraction

We seem to make use of structures of different sorts,

- some of different sizes at the same level of abstraction, i.e. AGGLOMERATION.
- others at different levels of abstraction i.e. using different ontologies, i.e. INTERPRETATION.

Various fragments are recognised in parallel and assembled into larger wholes which may trigger higher level fragments, or redirect processing at lower levels to address ambiguities, etc.

Here we have some of the fragments at the level of configurations of dots, and the next abstraction level, configurations of continuous line segments

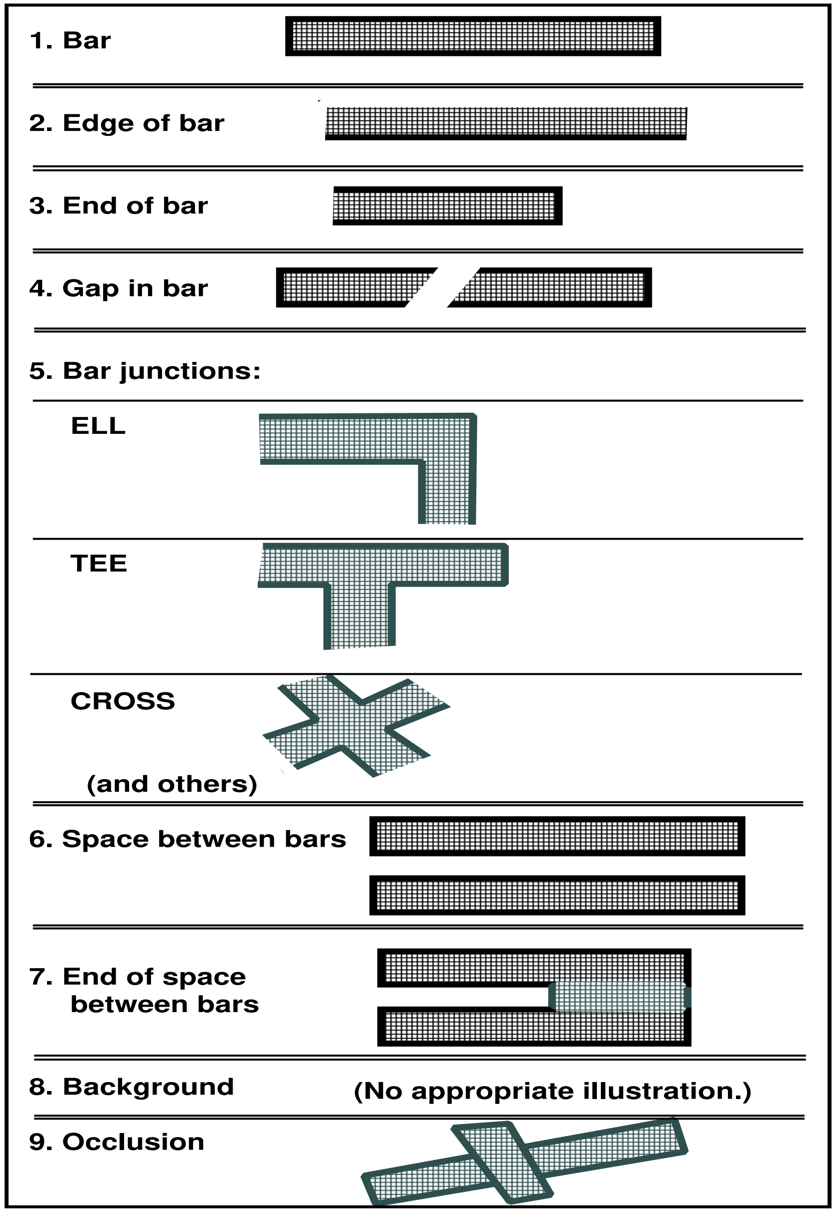


Useful fragments at one level of abstraction

Here are some of the significant fragments detectable in the domain of overlapping laminas made from merged rectangular laminas.

These might be worth learning as useful cues if the system can detect that they occur frequently.

Just one of MANY kinds of perceptual learning – linking newly developed perceptual mechanisms with extensions to ontologies and representations used

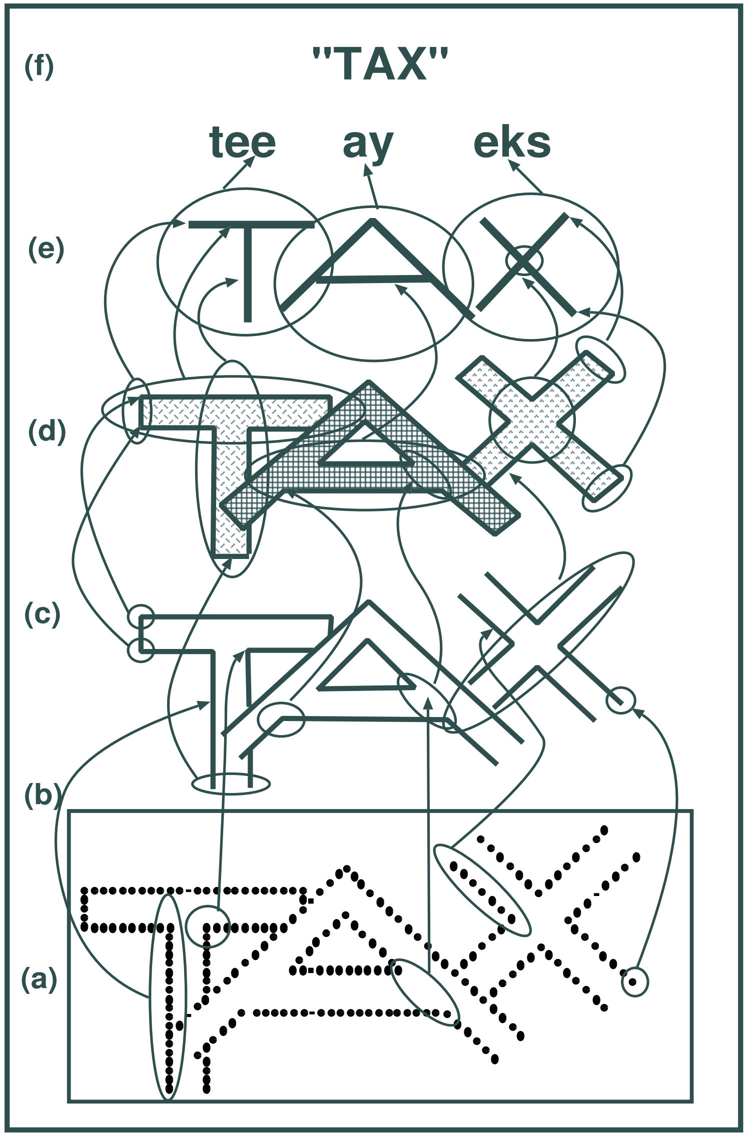


Aspects of perception: multiple levels of structure

Layers of interpretation of a 2-D dot pattern.

There are several ontologies involved, with different classes of structures, and mappings between them.

- **At the lowest level the ontology may include dots, dot clusters, relations between dots, relations between clusters. All larger structures are agglomerations of simpler structures.**
- **Higher levels are more abstract – besides grouping there is also interpretation, i.e. mapping to a new ontology.**
- **Reading text would involve even more layers of abstraction: mapping to morphology, syntax, semantics, world knowledge**



From *The Computer Revolution in Philosophy* (1978)
http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html

Multi-layer perceptual architectures

CONJECTURE: concurrent perceptual processing occurs at many different levels of abstraction.

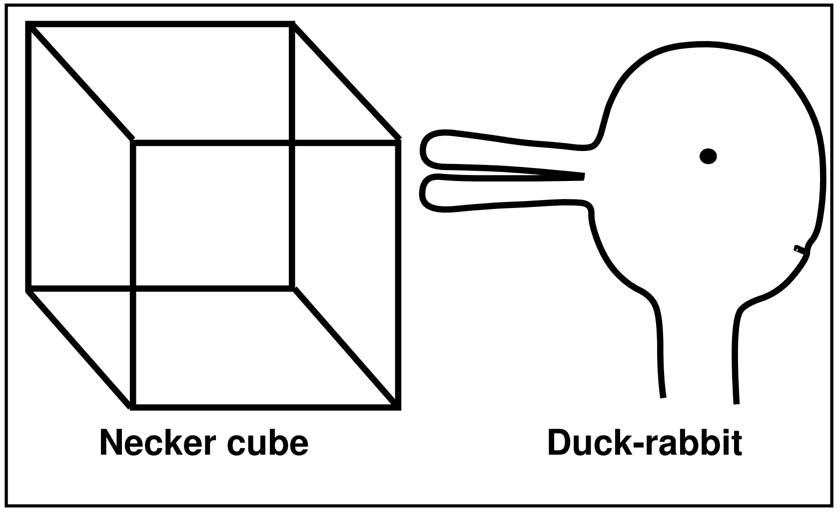
- Sub-systems at different levels can interact with other sub-systems, including interrupting them by providing relevant new information or redirecting "attention" or altering thresholds.
- Sometimes a higher level subsystem (e.g. word recogniser) will reach a decision before lower levels had finished processing.
- Sometimes a speedy high level decision will be wrong!

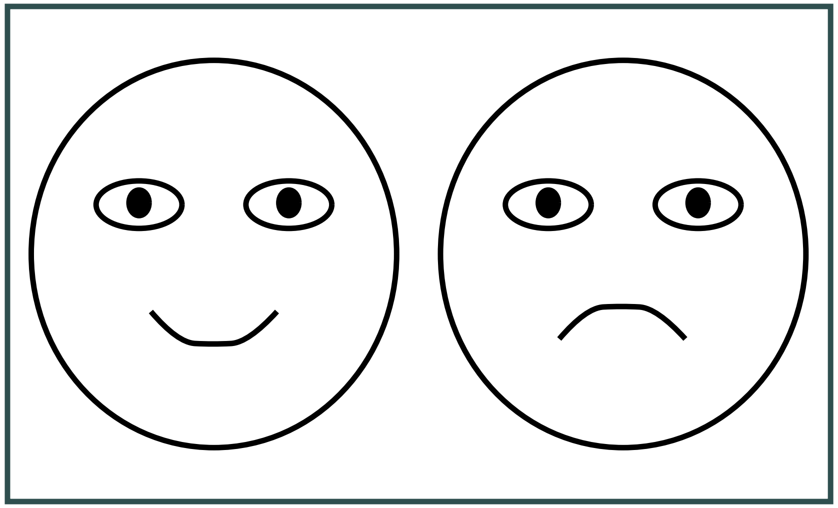
  It can be corrected later as information flows upwards.
- Different low level styles (dots, dashes, colour-boundaries) can feed into the same higher level domains: hence different styles of depiction may receive the same high level interpretation.

The different perceptual sub-systems will have different sorts of relationships to central architectural mechanisms. E.g. some may control reactive behaviours, while others provide information for learning associations relevant to deliberative layers.

But there's more than different classes of physical and geometrical properties to be perceived.

Ontologies 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...

We need to investigate the kinds of ontologies that visual mechanisms use, whether produced by evolution, development, or learning.

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. The justification for assuming there are other minds was increased biological fitness rather than evidence available to individuals.)

Affordances

Perceived affordances are even more subtle and complex than perceived non-physical states like mental states of other agents:

- **They essentially involve what does not exist, but might exist:**
- **Opportunities, constraints, threats, obstacles, etc.**
- **The ontology of affordances is defined relative to the goals, needs, capabilities of the perceiver: an important kind of subjectivity in perception.**
- **The ontology is inherently modal insofar as it is concerned with what can or cannot happen and the constraints, implications, etc. related to various possibilities.** (I can lift the mug but not using third and fourth fingers of one hand).
- **Some affordances involve rapid evaluations that trigger "alarm" mechanisms to interrupt, modulate, or redirect processing. This corresponds to a class of emotions, which vary according to which portions of the architecture are affected.**
- **E.g. tertiary emotions (characteristic of humans) involve disturbances of control (e.g. of attention) in the meta-management layer.**

No time for full discussion of affordances. See talks on vision
http://www.cs.bham.ac.uk/research/cogaff/talks/

More on the CogAff schema

The CogAff schema defines a variety of types of components of information-processing architectures, with functions determined by different boxes in the grid, and allows various possible types of information linkages, which may or may not be present in different instances.

To that extent **CogAff subsumes a wide variety of possible architectures**.

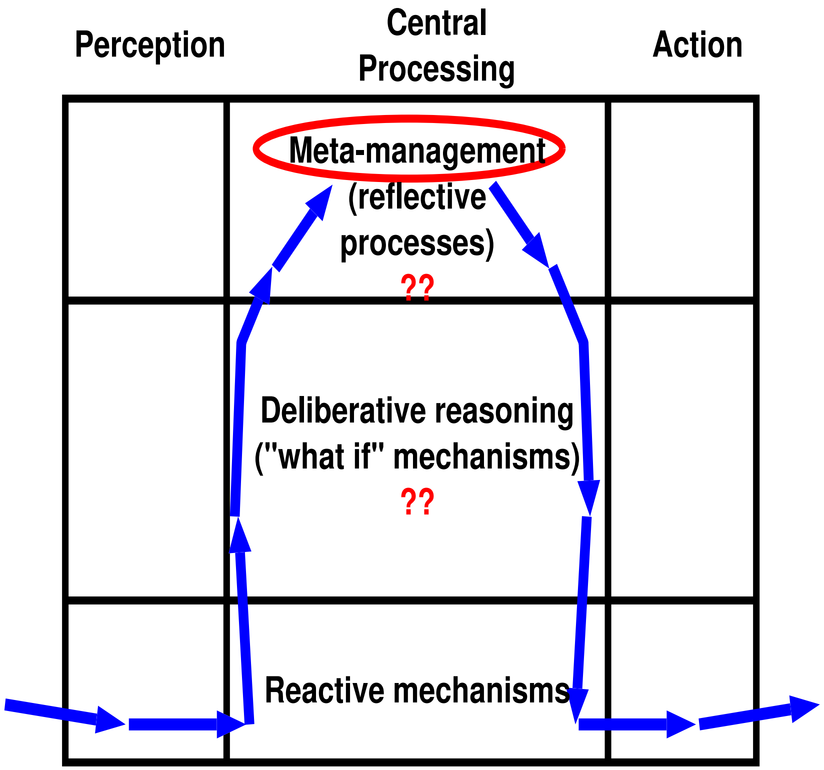
It does NOT specify how control flow must happen, or dominance of control: many options are left open.

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

Instead of saying "my architecture is best" theorists should show how their architectures are like and unlike other possible architectures allowed by the scheme, and what the trade-offs are, relative to various niches (possible sets of requirements).

We give a few examples of architectures consistent with Cogaff.

The “Omega” model of information flow



Rejects layered concurrent perceptual and action towers separate from but linked to central tower. (Only “peephole perception”?)

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

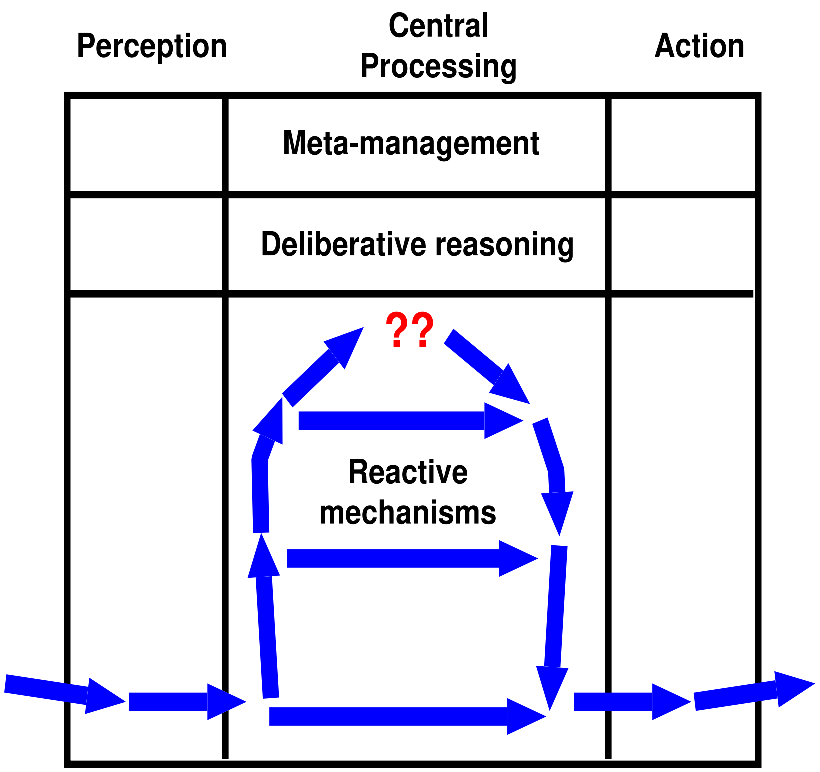
Some authors propose a “will” at the top of the omega.

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.

No multi-window perception or action?

Some supporters deny that humans use deliberative mechanisms, even to get to overseas conferences to give talks about subsumption architectures.



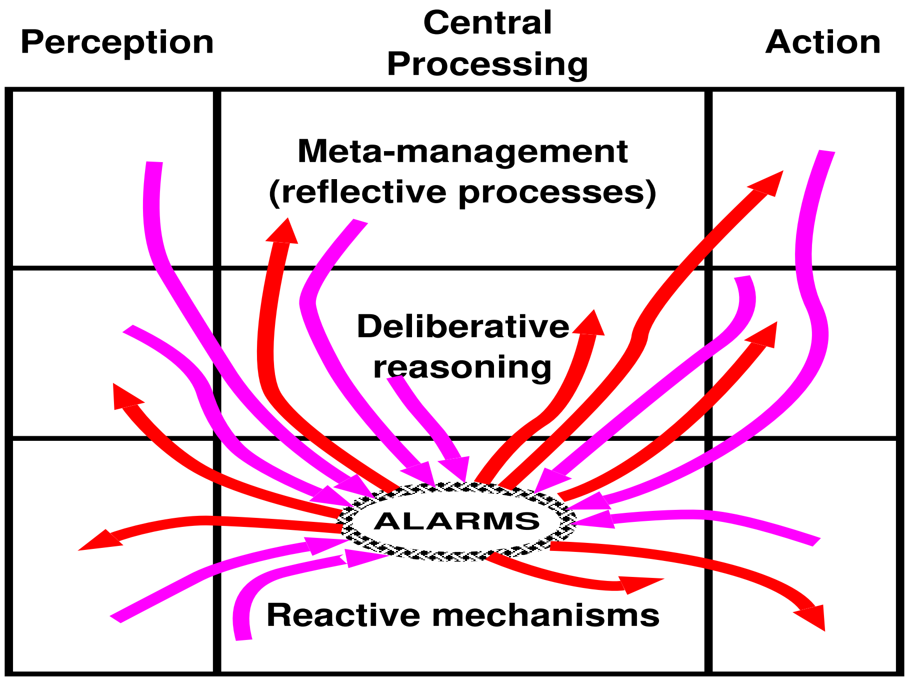
Brooks now writes about deliberative mechanisms:
http://www.dsic-web.net/meetings/oy8guwod/papers.html

As processing grows more sophisticated, so it can become 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!

Many variants are possible. E.g. purely innate, or trainable.
E.g. one alarm system or several?
(Brain stem, limbic system, ...???)



Many different kinds of emotional states can be based on such an alarm system, depending on what else is in the architecture.

Dont confuse the alarms (and emotions they produce) with the evaluations that trigger them.

See Cogaff papers and talks

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

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

Emotions and control mechanisms

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? Can it have emotions?

Evaluations are another bag of worms.

Architecture-based conceptions of emotion

A first crude subdivision:

- Primary emotions - initiated in the reactive layer
- Secondary emotions - initiated in deliberative layer
- Tertiary emotions - initiated in meta-management layer

Many further sub-divisions.

- **Reflexes:** "straight-through" connections in the reactive layer.
- **Second-order emotions:** triggered by the fact of having an emotion, e.g. guilt at feeling jealous.
- **Perturbances:** disturbances of control of attention in the meta-management layer.

Most theories of emotion fail to do justice to the fine structure of processes and mechanisms and the variety of cases.

- There's too much focus on external manifestations, and physiological processes.
- There are fallacious arguments that intelligence needs emotion.
  (e.g. Damasio)

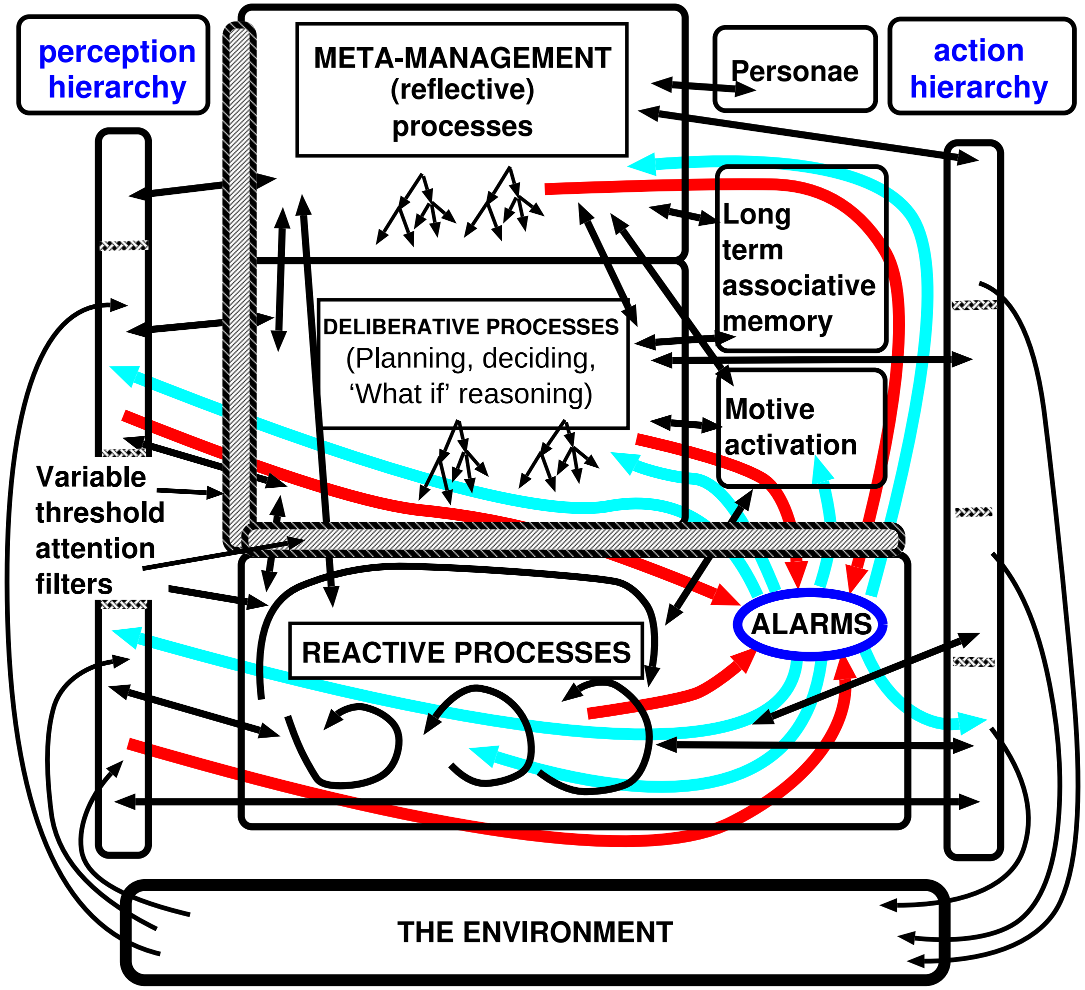
See Cognition and Affect project papers and talks

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

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

The architecture of a human mind

(first draft – see http://www.cs.bham.ac.uk/research/cogaff/)



See other cogaff papers and talks for details

Our informal concept of “emotion” is very confused.

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, long term grief
- **Does an emotion have to have some externally observable/measurable physiological manifestation when present?**
- **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!

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, or preserve a match.

– 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.

Hume: Reason is and ought to be the slave of the passions.

Roughly: without affect there is no reason to do anything.

Affect: whatever initiates, preserves, prevents, selects between, modulates, actions.

Some affective states are derivative on others (e.g. wanting X because it is conducive to, or prevents or preserves Y, etc.)

Direction of fit

This is related to what some philosophers have called the difference in "direction of fit" between beliefs and desires. We generalise this to larger classes of information states.

- **Belief-like states**
- **Desire-like states – Desire-like states may be explicit or implicit.**

There are some states which are neither belief-like nor desire-like, insofar as they merely represent possibilities which are being thought about or considered as options.

- **Contemplative states**

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 – e.g. reflexes don't seem to involve belief-like or desire-like states (from the point of view of the organism)

- **Pure reactive tendencies (e.g. eye-blink reflex).**

Are these "affective" states? There's no *right* definition of such a vague notion.

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.

People who require emotions to be “episodic” i.e. actively affecting some process, are unable to allow for long term emotions that can be temporarily dormant, e.g. grief, jealousy, obsessive ambition.

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.)

NB: Not all emotions are useful: many are seriously dysfunctional, e.g. obsessive jealousy, destructive anger.

Any theory that treats all emotions as useful does not do justice to the richness of human mentality.

Often emotional reactions are useful for immature individuals who cannot yet understand the pros and cons of alternative actions – they may be trained, or influenced by genes, to react emotionally, e.g. to avoid harm. Later when they are more mature they learn to be more in control, and less emotional.

Notes on varieties of learning

All this implies that many more kinds of learning and development are possible than normally allowed for

- Which new mechanisms develop, or change
- Which new links develop
- Which new formalisms develop in various subsystems
- Which new ontologies develop or are extended in various subsystems
- Which new uses for old mechanisms and capabilities develop
- Many kinds of perceptual learning: e.g. learning to perceive new mental states
- Learning new hierarchically organised actions (e.g. athletic skills, musical skills, dancing skills, speaking skills, tree climbing, nest building....)
- How processes are made faster, smoother (various ways)
- How processes are made more robust, reliable,
- How limitations are discovered and evaluated
- Development of new sub-personalities, for different contexts
- How new goals, desires, preferences, priniciples, values, are learnt
- Many kinds of social learning e.g. linguistic, ethical, and which kind of personality to use in various contexts

It follows that *defining* "architecture" as something that does not change closes off research into self-modifying architectures: e.g. human minds between birth and adulthood.

Learning and Affect

Affect can be involved in the production of learning: e.g. learning to avoid things whose consequences are evaluated as bad or unpleasant, and learning to do things whose consequences are evaluated positively.

However there are kinds of learning that change what is desired, enjoyed, liked, disliked, thought to be good etc.

I.e. learning can change the mechanisms involved in affect, so that affective reactions change.

Sometimes this can be harmful: e.g. learning to like smoking, or too much alcohol and other drugs.

(Biological architectures are so complex that many things can go wrong. All complex software systems are prone to bugs, including software systems produced by evolution, e.g. nationalism, religion...)

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, robots...).

3. Finding out which architectures are relevant and how to describe them well is a hard research problem. Designs need to be related to niches, and trade-offs understood, not fitness measures.

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 scientific and philosophical 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 claim that facts about frontal lobe damage imply that intelligence *requires* emotions.

9. We need toolkits that are suitable for exploring a wide variety of types of architectures in order better to understand the possibilities and trade-offs.

Our SimAgent toolkit does that
http://www.cs.bham.ac.uk/˜axs/cogaff/simagent.html

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.

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)

The remaining slides are selected from other talks and could be used if this presentation is allowed to extend over several hours!
They are not well organised.

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.

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.

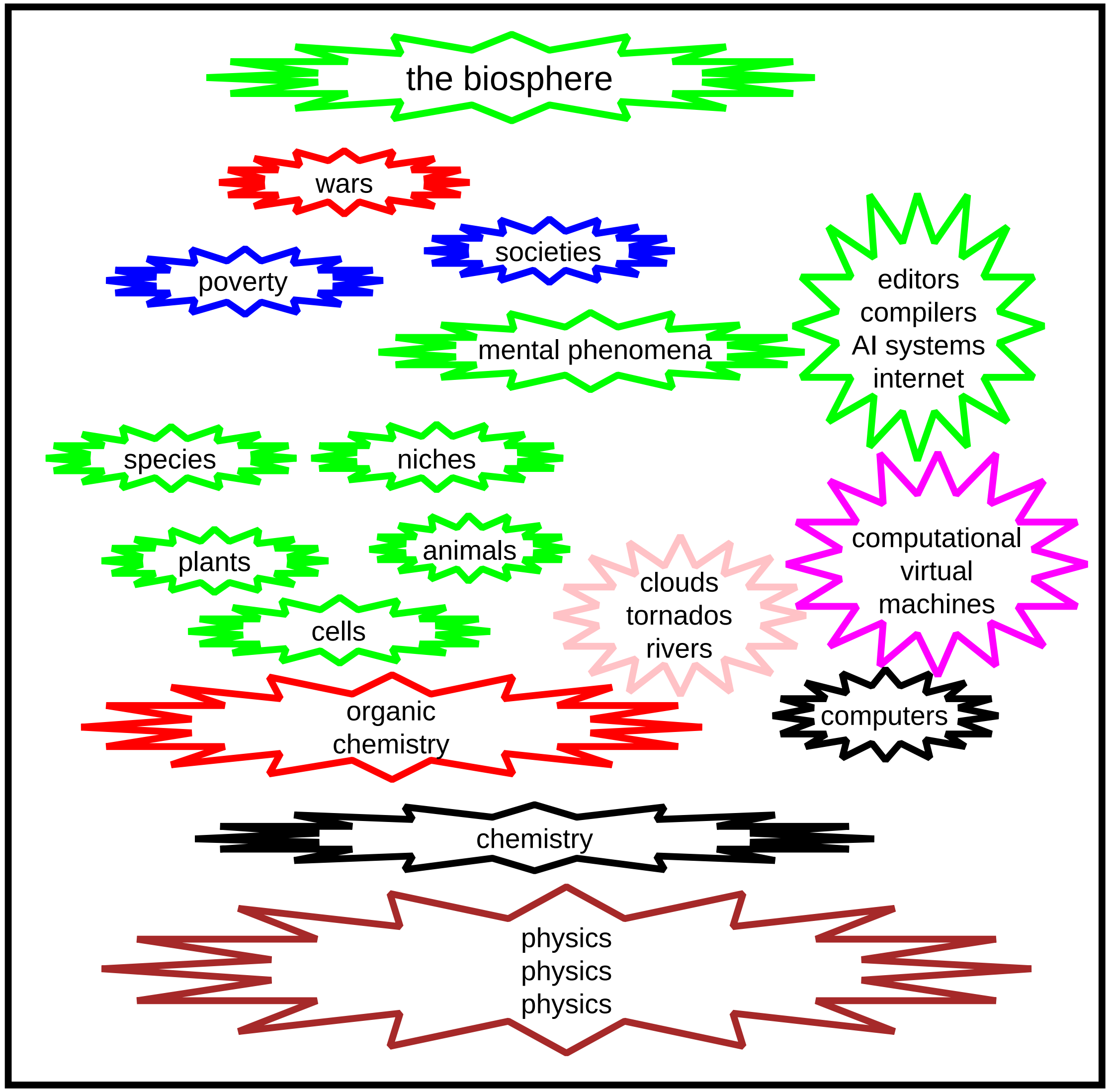
This leads to many philosophical problems about supervenience, implementation, etc.

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.

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.

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 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.