Presented at Birmingham Cafe Scientifique, 2004, and
NWO Cognition Programme
Utrecht 24 Jun 2005 http://www.nwo.nl/

DO INTELLIGENT MACHINES, NATURAL OR ARTIFICIAL, REALLY NEED EMOTIONS?

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

The (briefly fashionable?) belief that emotions are required for intelligence was mostly based on wishful thinking and a failure adequately to analyse the variety of types of affective states and processes that can arise in different sorts of architectures produced by biological evolution or required for artificial systems.

This work is partly a development of ideas presented by Herbert Simon in the 1960s in his 'Motivational and emotional controls of cognition'. (Simon, 1967)

Online at
http://www.cs.bham.ac.uk/research/cogaff/talks/#cafe04
http://www.slideshare.net/asloman

Last updated (January 16, 2014)

What is Life?

Living things are deeply intertwined, constantly interacting, mixtures of

- matter
- energy
- information (about many different things, inside and outside organisms)

NOTE: I use "information" in a sense that is closer to Jane Austen than Shannon:
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/austen-info.html`

I suspect that if Alan Turing had lived longer he might have combined his ideas about information processing in digital computers (Turing, 1936) and his ideas about continuous and discrete interactions in chemical systems (Turing, 1952) to provide a new, richer foundation for theories of evolution by natural selection.

We need a theory explaining how the mechanisms, structures and processes involved in evolution produce new mechanisms, structures and processes that repeatedly enrich both evolution and its products, by repeatedly extending the kinds of information used and the designs for information-processing mechanisms: the Meta-Morphogenesis project.

`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html`

Doing that will provide new richer, deeper, ways of thinking about designs for minds of many kinds, including microbes, mice, monkeys, humans, and perhaps future machines.

This presentation is about some of what needs to be explained, including affective (e.g. emotional) mechanisms, states and processes – aspects of control functions of minds.

This extends Herbert Simon's ideas: (Simon, 1967; Sloman, 1987, 1982)

Compare: Tibor Ganti, *The Principles of Life*, (Ganti, 2003).

The Birmingham Cognition and Affect Project: `http://tinyurl.com/BhamCog#overview`

Key ideas

- Our ordinary concepts such as ‘emotion’, ‘consciousness’, ‘feeling’ involve too much muddle, confusion and context-sensitivity to be useful in posing scientific questions or formulating explanatory theories, except in the early stages of science.
- Those ideas evolved for purposes of ordinary communication among lay people and are fine for their original purposes but they subvert scientific communication and theorising (like trying to base physics on earth, air, fire, and water).
- They also diverge across cultures: (Wierzbicka, 1992)
- We can refine, extend, and subdivide – to produce new, more precise, more theoretically-based, concepts – if we analyse the kinds of states and processes that can occur in biological information-processing architectures.

  (Physics and chemistry revised and extended our concepts of kinds of matter and kinds of physical and chemical states and processes, as evolutionary theory did to our concepts of biological species).
- Different architectures can support different sorts of states and processes.
- We need to understand a wide range of architectures – not just humans:

  E.g. can an insect have emotions? (whatever we think emotions are).
  Or an octopus? Read what the keepers say about Otto the octopus here
  `http://www.telegraph.co.uk/news/newstopics/howaboutthat/3328480/Otto-the-octopus-wrecks-havoc.html`
- We should resist ‘wishful thinking’ when we try to do science.

We need to learn to think about information-processing architectures: how they evolve over millennia and develop within individuals in seconds, weeks, years...

What is an architecture?

A house, a ship a symphony, a computer operating system, a company, a novel, a poem, a mathematical proof, an organisation ... can have an architecture: What is it they have?

- Each of those entities is something complex with parts, which can also have parts that have parts that ....
- The parts can have various sorts of relationships to other parts, including being close to or remote from them, having influences in either or both directions, sharing resources, cooperating to perform some function, interacting with external objects, ...
- Some of the parts are physical, like the parts of a house or a ship, while others are more abstract such as the overture of a symphony or the mission of the organisation.
- Talking about the architecture of X is talking about what the (physical and non-physical) components of X are, how they are related to X and to one another, what they do (including changing the architecture) why they do it and how they do it.
- An architecture can change, e.g. acquiring new parts, new connections, new functions.
- Biological architectures grow themselves! (Chappell & Sloman, 2007)

In that sense a mind can have an architecture too: it has components that perform various functions, including perceiving, “digesting” information, making inferences, learning, storing information, generating motives, forming plans, forming theories, selecting motives to act on, controlling execution of plans, detecting inconsistencies, resolving conflicts, detecting signs of danger or signs of useful opportunities, and many more.

The need for a specialised ontology

- The concepts used in human languages in various cultures are products of millions of years of biological evolution and shorter periods of cultural evolution, both influenced by physical and biological features of the environment though not necessarily fully determined by them.
- These concepts and the grammatical forms in which they are combined, along with social practices, technologies and individual skills are an enormously rich resource, finely tuned to meeting many practical and convivial needs of the communities that use them.
- But such concepts often have to be extended, modified, and sometimes rejected, for purposes of scientific theorising and technological advance – understanding what sorts of things exist, what sorts of things can exist, and what the constraints on various possible structures, events and processes are, and understanding how to make new types. (Wierzbicka, 1992).
- Examples of scientific advance with profoundly important explanatory power, predictive power and practical uses include (a) Newton's and Einstein's modifications of colloquial concepts of space, motion, force, and causal interaction and (b) replacement of early concepts of "kinds of matter" with the theory of physical elements and molecular compounds.
- Another example was replacement of observation-based concepts for classifying living things with a system based on evolutionary as well as structural and behavioural relationships, which relocated whales among mammals rather than fish.
- Similar conceptual reorganisation is required for a science of mind, but many scientists in that area (unwittingly?) restrict their thinking to familiar sets of concepts, including culture-specific concepts.
- This is especially true of psychology and other disciplines studying human minds, attempting to use either colloquial, or observation-based concepts of knowledge, belief, emotion, reasoning, whereas scientific advance requires deeper concepts and theories, often conflicting with colloquial versions.
- I'll try to show how the best (future) alternatives will need theories about information processing architectures and mechanisms of various sorts of evolved and constantly developing minds.

The Design-based Approach to study of mind

When scientists discuss experimental observations, and evaluate theories, they often formulate questions using language that evolved for informal discourse among people engaged in every day social interaction, like this:

What does the infant/child/adult/chimp/crow (etc) perceive/understand/learn/intend (etc)?
What is he/she/it conscious of?
What does he/she/it experience/enjoy/desire?
What is he/she/it attending to?
What is he/she/it trying to do?

I suggest that although those questions are very useful in everyday life and some clinical settings, if we wish to gain increased scientific insight we should also ask questions like:

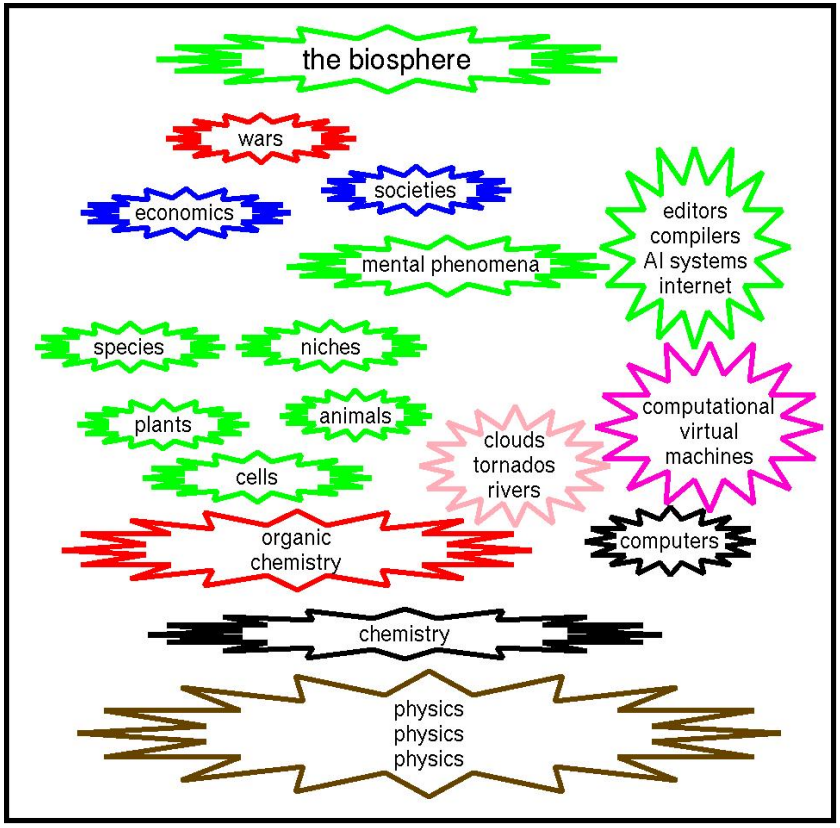
Which parts of the architecture are involved?
What are their functions?
What kinds of information do they acquire and use?
How do they do this?
What is the architecture in which they function?
How is the information represented? (It could be represented differently in different subsystems.)
When is the information used?
What kinds of manipulations and uses of the information occur?
What mechanisms make those processes possible?
How are the internal and external behaviours selected/controlled/modulated/coordinated?
How many different virtual machine levels are involved and how are they related (e.g. physical, chemical, neural, subsymbolic, symbolic, cognitive, social ...)?

In other words: The scientific study of mind needs to be “design-based”.

(That does not mean based merely on physics, chemistry and physiology.)

Architectures on different scales and levels

Our world has vast numbers of richly interacting, constantly changing, architectures at different levels of scale and levels of abstraction, some of them products of biological evolution, and some of them products of products, e.g. man-made machines, human socio-economic systems, battles, wars,...



There remain many puzzles at all levels, including puzzles about the nature of the physical world.

Uses of information

All organisms use information, but there are enormous variations in:

type of information, where it is obtained, how it is represented/encoded, how it is manipulated, what it is used for, what mechanisms are used, whether it is used by the whole organism (location of food) or only a part (repair of damaged tissue), etc.

Some major sub-categories (related to emotions in different ways):

- Factual, or potentially factual: Information about an object, state of affairs, or process that exists, existed, will exist, or doesn't exist, but could exist.

  Includes theoretical information about what could exist and under what conditions.
- Control information, about what to do, starting something, modulating a process, stopping a process, re-starting, storing information, selecting between competing alternatives ... (includes goals, intentions, preferences, plans).
- Evaluative information, about what's desired or desirable, preferred, valued, undesirable, rejected,

Notes:

David Hume noticed that for an agent factual information is useless without the others.

The simplest organisms may use only control information, or control and evaluation information.

Different forms of encoding are useful for different sorts of information in different contexts.

Different organisms, or different subsystems in an organism may use different information, and different forms of representation.

Primary uses of information are internal, for controlling actions.

Later information, or new forms of information can be 'parachuted' in, if required, either during evolution, or during learning development, or when reasoning, deciding.

Information use in organisms

All organisms need control information, otherwise nothing will happen.

- All the control information may be genetically provided (viruses?), or may be schematic, with slots filled by conditions after birth/hatching.
- Meta-control information may be needed for turning subsystems on or off, or dealing with conflicts, if two incompatible actions are selected.
- Some organisms develop their own control information, either by learning (e.g. reinforcement learning can build up associative control subsystems), or by instantiating genetic schemas, or by other means.
- Evaluative information (tastes, preferences) can combine with other information to generate control information: e.g. negative evaluation of a perceived situation can trigger a goal to change the situation, which can interact with planning mechanisms to produce a plan, which can trigger evaluations, that cause the plan to be changed.
- Similar evaluations can be triggered by action, or plan execution.

Within a system with such mechanisms, different sorts of affect, including desires, preferences, emotions, moods, attitudes, values etc. can have various functional roles.

Some forms of affect are possible only if the factual information contents (see previous slide) include meta-cognitive information about what was thought, perceived, decided, etc., and how it relates to goals, preferences, values, etc. (Sloman, 1978, Chap 6).

We refer to this as “Meta-management”, following (Beaudoin, 1994).

Some of the possible forms of processing involving goals (or motives) are specified in (Beaudoin & Sloman, 1993), e.g. detecting conflicts, reacting to near-misses, re-ordering goals, etc.

Two Caveats regarding "design" and "information"

Talking about a **design** does not presuppose the existence of a **designer:** biological evolution produced designs of many kinds, performing many functions, including reproduction, locomotion, manipulation, perception, learning, ...

The word "information" is not used here in the sense of Claude Shannon, but in the much older, everyday sense in which Jane Austen used it, e.g. in *Pride and Prejudice*, where information is often acquired, stored, and used to achieve things.

For examples of her use see `http://tinyurl.com/CogMisc/austen-info.html`

That's the main sense in which "information" is relevant to psychology and biology, though it needs to be extended, and Shannon's sense is relevant in other contexts.

The Meta-Morphogenesis project combines early and late themes in Turing's work, investigates transitions in types of information-processing and mechanisms of information- processing produced by biological evolution, including changes in mechanisms of evolution, and changes in learning mechanisms during development. See:

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

NB:
chemical information processing came first, and is still crucial for living things, e.g. building brains.

Example: A simple (insect-like) architecture

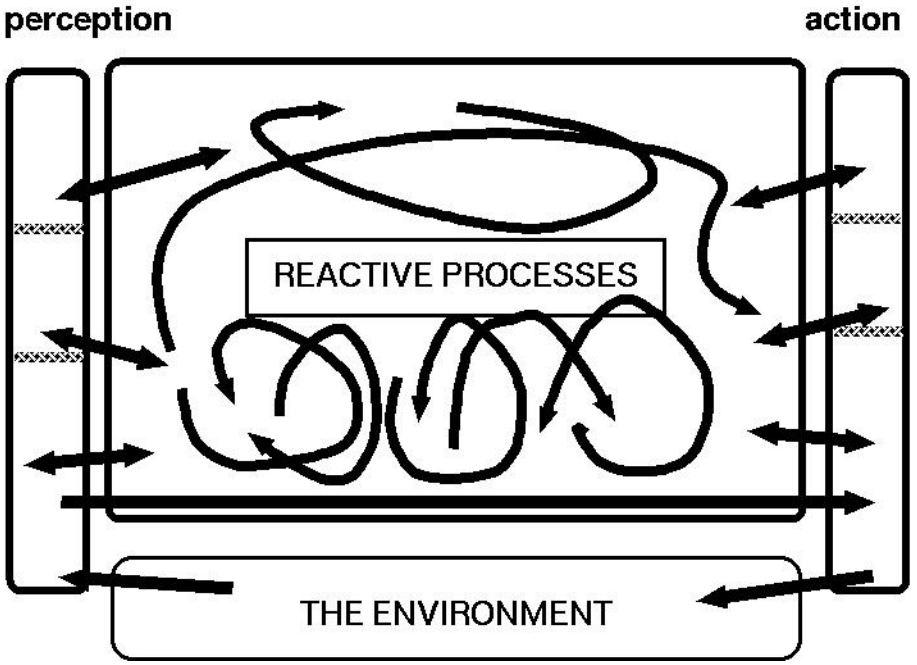
A reactive system does not construct complex descriptions of possible futures, evaluate them and then choose one.

It simply reacts: internally or externally, or both.

Several reactive sub-mechanisms may operate in parallel.

The arrows loosely indicate types of information flow (including control information).

Processing may use a mixture of analog (continuously changing) and discrete mechanisms (changing step-wise).



An adaptive system with reactive mechanisms can be a very successful biological machine.

Some purely (or mostly?) reactive species also have a social architecture, e.g. ants, termites, and other insects. Also microbes communicating chemically.

Most animals are purely reactive

The vast majority of species are entirely reactive, lacking deliberative capabilities (explained later).

If an ant could consider several possible actions, and for each one imagine several steps ahead, evaluating the options, and then choose one of the multi-step options and carry it out, it would not be purely reactive – It would include deliberative capabilities.

Most organisms (e.g. microbes, invertebrates, including insects) have more or less complex architectures, including perceptual and motor processes functioning at different levels of abstraction, e.g. detecting moisture, detecting an opportunity to mate.

But they don't appear to explore future possibilities two or more steps ahead in their minds before deciding what to do.

If they did they would have deliberative competences: they would not be purely reactive.

- Purely reactive biological species that are precocial may have very many genetically determined capabilities, possibly modified by environmentally driven learning.
- They can be biologically very successful: e.g. most of the biomass on this planet.
- Rapid creative learning requires additional mechanisms, not discussed here.

A more detailed exposition would show that there is not a simple dichotomy between reactive and deliberative. (Sloman, 2006)

- E.g., there are "proto-deliberative" systems, in which two or more alternatives for a "next step" compete, and a whole variety of intermediate cases, with "fully deliberative" systems as an extreme.
- Some reactive species collaborate deliberatively using pheromone trails to record explorations.

More on architectures: `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604`

What does all that have to do with emotions?

CLAIM: "Emotional" is a highly polymorphic concept, with diverse sub-types.

The most general concept that fits our fuzzy and indeterminate mish-mash of uses of the words 'emotional' (the more basic word) and 'emotion' is:

A state in which a monitoring mechanism acquires a tendency (i.e. a disposition, possibly suppressed) to abort, redirect, or modulate some other process or collection of processes. (Some people consider only the realised dispositions to be emotions.)

Example: a house-fly consuming food detects something rapidly descending towards it: the 'alarm' mechanism aborts eating and triggers escape behaviour.

States and processes that people label as 'emotions' vary enormously, involving both evolutionarily old and new mechanisms, producing both short term and long term states (e.g. grief, jealousy, infatuation, ambition) – some of which may occur in future robots.

There need not be any specific bodily changes involved: the important things are control changes (some in virtual machines)

E.g. a mathematician working on a new proof notices a fallacy caused by implicit division by zero. The resulting emotion triggers a disposition to switch to investigating the offending step in the proof.

Some of these disruptions can be unconscious – like the people who are jealous or infatuated and don't realise it, though it is evident to their friends (a phenomenon exploited by novelists and playwrights).

Often the tendency or disposition is not resisted and has immediate effects, but not always.

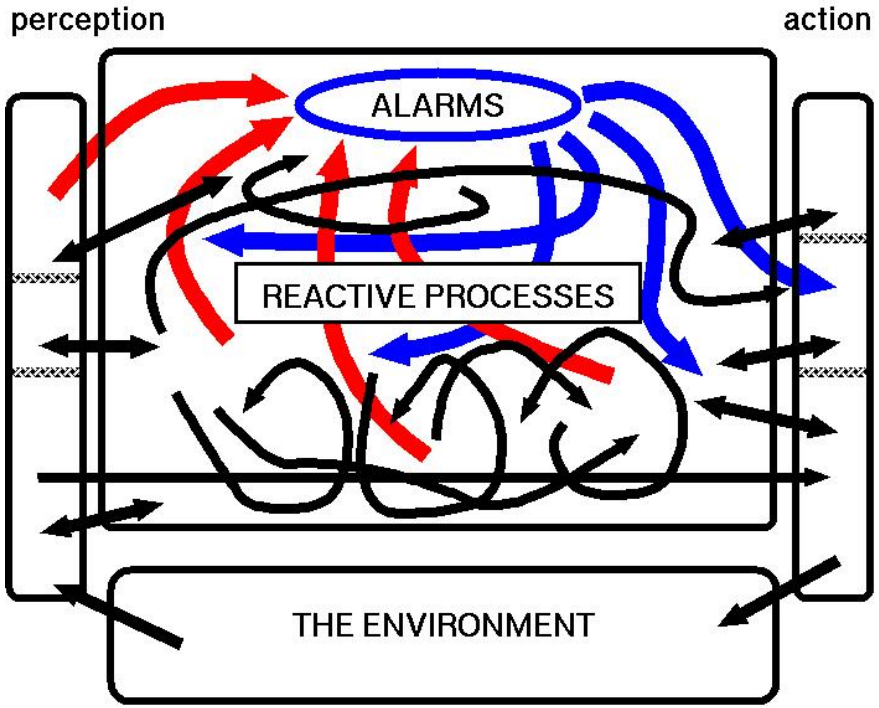
More subtly, the disruptive tendency may be suppressed or overridden, but persist, competing for control, and sometimes taking over (e.g. a grieving parent reminded of a dead child months or years after the death, discussed in more detail in (Wright, Sloman, & Beaudoin, 1996)).

Primary emotions in insects?

Even insects may need simplified 'emotions':
e.g. detecting and reacting to unexpected dangers or opportunities, using fast pattern recognition mechanisms.

'Alarm' mechanisms running in parallel with other things, using fast pattern recognition, can detect events, opportunities, threats, or new information sources, that indicate an immediate need to trigger some interference with 'normal' processes, such as

- aborting
- accelerating
- decelerating
- redirecting
- freezing,....



Humans, and many other species, need far more complex architectures.

NOTE: Much empirical research reveals only shallow behavioural consequences of deep mechanisms and architectures, and study shallow verbal classifications, e.g. ignoring deep cognitive and affective reactions to processes in other people. Good novelists, playwrights, and poets sometimes have richer insights, as Keith Oatley has often pointed out, though they lack concepts required for deep explanations.

Related points are made in (Donald, 2002) (spoilt initially by exaggerated rants against "reductionists"!).

Proto-deliberative and deliberative mechanisms

Consider a reactive organism O extended with a decision mechanism D that receives and compares inputs from two sensors SF (detecting amount of food nearby) and SP (detecting proximity of a predator), each with an associated threshold value, TFood and TPredator.

- If SF does not exceed TF and SP does not exceed TP then D does nothing.
- If only one of SF or SP has exceeded its threshold, then D simply passes on the signal from the sensor to the motor mechanisms to produce food-seeking or predator-avoiding behaviour, as if it were a purely reactive mechanism.
- If both have exceeded their thresholds, then D passes on the signal from the one with the greater excess, suppressing the other — i.e. it is a 'winner-takes-all' mechanism.
  (This assumes the scales for SF and TF are somehow calibrated as commensurable!)

We can describe this as a 'proto-deliberative' mechanism (Sloman, 2006).

It 'considers' two options and selects one.
(Unfortunately, Arbib calls this 'deliberative' without qualification in (M. A. Arbib, 2002).)

In contrast fully deliberative systems, can explore and compare multi-stage branching futures, as AI planning mechanisms and human planners do.

With meta-deliberative capabilities O can deliberate about how to deliberate: e.g. "Is more information needed?" "Can the decision be postponed?" etc...

If the comparison process takes some time, we can say that while the comparison has not yet been completed, the organism O is in a 'state of indecision' of a rather primitive sort.
(This does not require O to be aware of being undecided: it may lack meta-cognitive capabilities.)

Actual or potential disturbances (perturbant states) involving deliberative mechanisms an produce "secondary emotions", based on higher cortical functions. (LeDoux, 1996)

Secondary and tertiary emotions

There are various ways in which reactive architectures can be extended with additional functionality.

Biological evolution seems to have explored them in great detail, with a huge range of variation across many forms of life, the vast majority of which are only reactive, and lack both deliberative capabilities and meta-cognitive capabilities used for explicit monitoring and control of information-processing mechanisms.

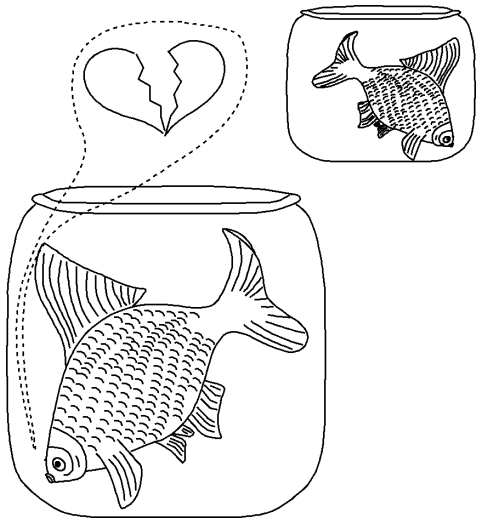
I don't think the vast majority of ordinary language words and phrases for describing mental states and processes have sufficient precision to serve the long term aims of science – as they are like labels for kinds of "stuff" available before the development of modern physics and chemistry.

Compare ways of grouping animals before modern evolution-based biology (e.g. treating whales as fish because of their shape, behaviour and habitat).

Developing deep explanatory theories about the information-processing mechanisms with varying kinds of complexity in organisms allows us gradually to replace ordinary language labels with a much richer, more precise theory-based terminology.

But that requires good architectural theories – including a theory of the "space of possible architectures" since different architectures will support different sorts of affective states and processes.

Can a goldfish long for its mother? If not why not? What's missing?



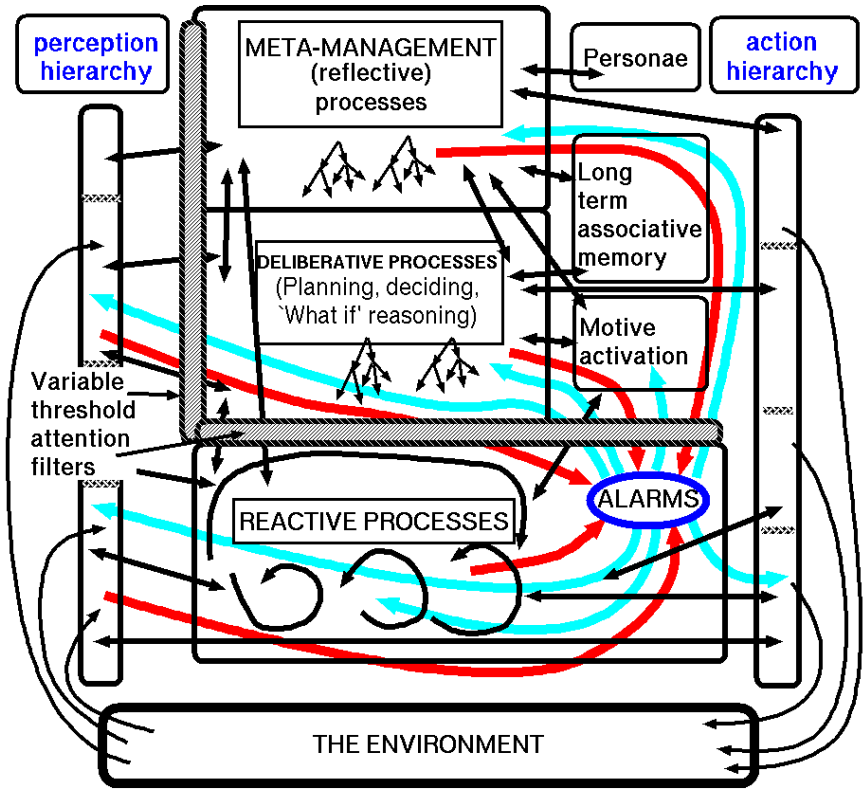
Something more human: H-Cogaff

In humans, more varied subsystems operate concurrently, performing different tasks.

Some are evolutionarily old 'purely reactive' subsystems, and similar to mechanisms in many other kinds of animals. (Bottom layer.)

Other subsystems (e.g. deliberative and meta-management layers) are newer and do tasks that far fewer animals can perform.
(E.g. thinking about past events, remote events, future events, hidden causes, and what another individual sees, thinks, wants, intends, etc.)

Perception and action operate concurrently at different levels of abstraction, in relation to different central sub-systems.



By considering "alarm" processes in the different layers we can distinguish more kinds of emotion-like states, e.g. primary, secondary, tertiary, and far more, depending on how they differ in detail.

The ideas behind this diagram are explained in more detail in
`http://tinyurl.com/CogMisc/vm-functionalism.html`
`http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200307` (Background)

Note: mechanisms involved in human language use, not shown here, pervade most parts of the system in adult humans.

Damasio's Error

- In 1994 Antonio Damasio, a well known neuroscientist, published *Descartes' Error*.

  He argued that emotions are needed for intelligence, and accused Descartes and many others of not grasping that.

- In 1996 Daniel Goleman published *Emotional Intelligence: Why It Can Matter More than IQ*, quoting Damasio with approval.

- Likewise Rosalind Picard a year later in her ground-breaking book *Affective Computing*.

- Since then there has been a flood of publications and projects echoing Damasio's claim, and many researchers in Artificial Intelligence have become convinced that emotions are essential for intelligence, so they are now producing many computer models containing a module called 'Emotion'.

- Before that, serious researchers had begun to argue that the study of emotions, and related phenomena, had wrongly been neglected in psychology, and cognitive science, but the claims were more moderate e.g. (Boden, 1972)

  E.g. a journal called *Cognition and Emotion* was started in 1987.
  Even I had a paper in it in the first year. But H.A. Simon's work was mostly ignored.

Alas, all that theorising was not based on a deep understanding of varieties of interacting information-processing mechanisms required to explain human phenomena.
It did not embrace 'the design-based approach' (defined above).

Damasio's examples

Damasio's argument was partly based on two examples:

- Phineas Gage: In 1848, an accidental explosion of a charge he had set blew his tamping iron through his head – destroying the left frontal part of his brain.

  "He lived, but having previously been a capable and efficient foreman, one with a well-balanced mind, and who was looked on as a shrewd smart business man, he was now fitful, irreverent, and grossly profane, showing little deference for his fellows. He was also impatient and obstinate, yet capricious and vacillating, unable to settle on any of the plans he devised for future action. His friends said he was No longer Gage."

  `http://www.deakin.edu.au/hbs/GAGEPAGE/Pgstory.htm`

  Christopher Green, however, informs me that most popular reports on Gage exaggerate the effects of his injury. See `http://www.nthposition.com/anoddkindoffame.php`

- Elliot, Damasio's patient ('Elliot' was not his real name.)
  Following a brain tumor and subsequent operation, Elliot suffered damage in the same general brain area as Gage (left frontal lobe).

  Like Gage, he experienced a great change in personality. Elliot had been a successful family man, and successful in business. After his operation he became impulsive and lacking in self-discipline. He could not decide between options where making the decision was important but both options were equally good. He perseverated on unimportant tasks while failing to recognize priorities. He had lost all his business acumen and ended up impoverished, even losing his wife and family. He could no longer hold a steady job. Yet he did well on standard IQ tests.

  `http://serendip.brynmawr.edu/bb/damasio/`

WHAT FOLLOWS FROM THIS?

Both patients appeared to retain high intelligence as measured by standard tests, but not as measured by their ability to behave sensibly.

Both had also lost certain kinds of emotional reactions.

WHAT FOLLOWS?

Damasio's argument

Here is the essence of the argument Damasio produced, which many people in many academic disciplines enthusiastically accepted as valid:

There are two factual premises from which a conclusion is drawn.

P1 Damage to frontal lobes impairs certain emotional capabilities

P2 Damage to frontal lobes impairs intelligence

C Those emotional capabilities are required for intelligence

IS THIS A VALID ARGUMENT?

The conclusion does not follow from the premises.
Whether the conclusion is true is a separate matter, discussed later.

In fairness, Damasio's book did not present the argument in quite such a bare fashion.

However, something like this argument is often approvingly attributed to him.

(An example is mentioned later.)

Compare this argument

We ‘prove’ that cars need functioning horns in order to start, using two premises on which to base the conclusion:

P1 Damage to a car battery stops the horn working

P2 Damage to a car battery prevents the car starting

C A functioning horn is required for the car to start

DOES C FOLLOW FROM P1 AND P2?

Why did readers not spot the fallacy?

A moment's thought should have reminded Damasio's readers that

- two capabilities A and B could presuppose some common mechanism M, so that
- damaging M would damage both A and B,
- without either of A or B being required for the other.

For instance, even if P1 and P2 are both true, you can damage the starter motor and leave the horn working, or damage the horn and leave the starter motor working!

NOTE:

I am ignoring two points, for the sake of illustration.

- Without a battery some cars can be started by rolling them downhill.
- There may be some cars which have separate batteries for starter motor and horn, in which case P1 and P2 would would both be false generalisations.
  For such cars the premisses would be inappropriate because the phrase 'the battery' presupposes that there is only one.

Why were so many people convinced?

Why are so many intelligent people convinced by Damasio's argument?

I first criticised Damasio's argument in two papers in 1998 and 1999:

A. Sloman, (1998) Damasio, Descartes, Alarms and Meta-management, in *Proceedings International Conference on Systems, Man, and Cybernetics (SMC98)*, San Diego, IEEE, pp. 2652–7,

Available online: `http://www.cs.bham.ac.uk/research/cogaff/0-INDEX96-99.html#36`

A. Sloman, (1999) Review of Affective Computing by R.W. Picard, 1997, in *The AI Magazine*, 20, 1, pp. 127–133

Available online: `http://www.cs.bham.ac.uk/research/cogaff/0-INDEX96-99.html#40`

I have never seen these criticisms of Damasio's arguments made by other authors.

My criticisms were repeated in several subsequent publications.
Nobody paid any attention to the criticism and even people who had read those papers continued to refer approvingly to Damasio's argument in their papers.
Some even quote me as agreeing with Damasio!

Very intelligent people keep falling for the argument.

WHY? What's wrong with emotion researchers?

E.g. Susan Blackmore, a highly intelligent and experienced researcher, did not notice the fallacy when she approvingly summarised Damasio's theories. See page 285 of her otherwise excellent book *Consciousness: An Introduction (2003)*.
She has informed me that she agrees that the argument reported approvingly is fallacious.

A sociological conjecture

The best explanation I can offer for the surprising fact that so many intelligent people are fooled by an obviously invalid argument is sociological:

they are part of a culture in which people want the conclusion to be true.

There seems to be a wide-spread (though not universal) feeling, even among many scientists and philosophers, that intelligence, rationality, critical analysis, problem-solving powers, are over-valued, and that they have defects that can be overcome by emotional mechanisms.

This leads people to like Damasio's conclusion. They want it to be true.

And this somehow causes them to accept as valid an argument for that conclusion, even though they would notice the flaw in a structurally similar argument for a different conclusion (e.g. the car horn example).[*]

A research community with too much wishful thinking does not advance science.

Instead of being wishful thinkers, scientists trying to understand the most complex information-processing system on the planet should learn how to think (some of the time) as designers of information-processing systems do.

---

[*] *This is a general phenomenon: consider how many people on both sides of the evolution/creation debate or both sides of the debate for and against computational theories of mind tend to accept bad arguments for their side.*

A personal note

I find it curious that highly intelligent AI researchers who read my papers concerned with emotions apparently interpret my words as saying

what they wish I had said.

For example:

- My mention of Damasio is taken as an endorsement of his arguments: people either don't read or don't understand what I write about his arguments (e.g. that such theories cannot account for enduring emotions such as long term grief).
- A paper I wrote with Monica Croucher in 1981, entitled 'Why robots will have emotions' is reported as if it had been 'Why robots should have emotions'.

  See `http://www.cs.bham.ac.uk/research/cogaff/81-95.html#36`
- An online presentation attributes to me the completely daft theory that

  In each reasoning cycle:

  A goal is selected as the focus of attention

  The one that generates the strongest emotions (Sloman)

  I found that presentation when browsing the section on emotions in the euCognition web site.

  `http://www.eucognition.org/affect_emotion_articles.htm`

We need to counter this wishful thinking, sloppy science, and sloppy reporting?

To be fair ....

In fact Damasio produced additional theoretical explanations of what is going on, so, in principle, even though the quoted argument is invalid, the conclusion might turn out to be true and explained by his theories.

However:

- His theory of emotions as based on 'somatic markers' is very closely related to the theory of William James, which regards emotions as a form of awareness of bodily changes. This sort of theory is incapable of accounting for the huge subset of socially important emotions in humans which involve rich semantic content which would not be expressible within somatic markers (e.g. admiring someone's courage while being jealous of his wealth) and emotions that endure over a long period of time while bodily states come and go (such as obsessive ambition, infatuation, or long term grief at the death of a loved one).
- The key assumption, shared by both Damasio and many others whose theories are different, is that all choices depend on emotions, and especially choices where there are conflicting motives. If that were true it would support a conclusion that emotions are needed for at least intelligent conflict resolution.
- Although I will not argue the point here, I think it is very obvious from the experience of many people (certainly my experience) that one can learn how to make decisions between conflicting motives in a totally calm, unemotional, even cold way simply on the basis of having preferences or having learnt principles that one assents to. Many practical skills require learning which option is likely to be better.

  A lot of social learning provides conflict resolution strategies for more subtle decisions: again without emotions having to be involved.
- A terminological decision to label all preferences, policies, and principles 'emotions' would trivialise Damasio's conclusion.

So, let's start again: what are emotions, and how do they work?

Does a Crow need emotions in order to be intelligent?

SHOW BETTY MAKING A HOOK

http://users.ox.ac.uk/~kgroup/tools/photos.shtml

See the videos

http://users.ox.ac.uk/~kgroup/tools/movies.shtml

There are many more reports of this research in many web sites:

http://www.google.co.uk/search?num=30&hl=en&q=betty+crow+hook+&btnG=Search&meta=

WARNING:

The BBC reporter in one of the reports misquotes the researchers as saying that 9 out of 10 female crows can solve the problem, when what the researchers actually said was that Betty solved the problem 9 out of 10 times.

Beware of reporters, even from the BBC.

Does being emotional help a child solve his problem?

SHOW A CHILD FAILING TO UNDERSTAND HOOKS AND GETTING EMOTIONAL

See the video:

`http://www.cs.bham.ac.uk/~axs/fig/child-hook-train-big.mpg` [11 MB]

Compare Emre Ugur's robot that does lots of exploration and learns, without rewards. (Ugur, 2010)

What changes when a child who cannot solve a problem later becomes able to solve it?

One possibility: being trained with rewards and punishment, like a circus animal.

Another possibility – learning (with or without emotions):

- to perceive new affordances
- to acquiring a richer ontology
- to use new forms of representation
- to use new procedures or algorithms for making use of the above
- to recognise their relevance to particular problems
- to think instead of getting emotional!

Why should having emotions be necessary for such learning?
Contrast merely being driven by genome-created motives to find things out.

In order to explain all this

We have to think about

- information
- representations
- mechanisms
- architectures

And a host of related questions

- How many different kinds of learners are there?
- What are the implications of the differences?
- How and why did the different sorts evolve?
- Which learning mechanisms themselves result from learning?
- What roles do other learners (especially more advanced ones) play?

In comparison, talking about emotions explains very little: for learning that is prompted or aided by emotions requires a vast amount of machinery that can also work without emotions, and often does work without emotions when highly intelligent people are coming to understand something new and complex.

Weak students may need many emotional props and motivators, however. e.g. “stars” awarded by teachers. (Perhaps that’s what makes them weak learners?)

Some old ways to study emotions

There are many ways to study emotions and other aspects of human minds:

- Reading plays, novels, poems will teach much about how people who have emotions, moods, attitudes, desires, etc. think and behave, and how others react to them — because many writers are very shrewd observers!
- Studying ethology will teach you something about how emotions and other mental phenomena vary among different animals.
- Studying psychology will add extra detail concerning what can be triggered or measured in laboratories, and what correlates with what in the experiments.
  (As opposed to real life.)
- Studying developmental psychology can teach you how the states and processes in infants differ from those in older children and adults.
- Studying neuroscience will teach you about the physiological brain mechanisms that help to produce and modulate mental states and processes.
- Studying therapy and counselling can teach you about ways in which things can go wrong and do harm, and some ways of helping people.
- Studying philosophy with a good teacher may help you discern muddle and confusion in attempts to say what emotions are and how they differ from other mental states and processes.

There's another way that complements those ways.

Another way to learn: do some engineering design

Suppose you had to design animals (including humans) or robots capable of living in various kinds of environments, including environments containing other intelligent systems.

What sorts of information-processing mechanisms, including control mechanisms, would you need to include in the design, and how could you fit all the various mechanisms together to produce all the required functionality, including:

- perceiving,
- learning,
- acquiring new motives,
- enjoying some activities and states and disliking others,
- selecting between conflicting motives,
- planning,
- reacting to dangers and opportunities,
- communicating in various ways
- reproducing, and so on...

If we combine this "design standpoint" with the previously listed ways to study mental phenomena, we can learn much about all sorts of mental processes: what they are, how they can vary, what they do, what produces them, whether they are essential or merely by-products of other things, how they can go wrong, etc.

The result could be both deep new insights about what we are, and important practical applications.

The design-based approach – too fragmented now

The design-based approach is not new: over the last half century, researchers in Computational Cognitive Science, and in Artificial Intelligence have been pursuing it.

- Because the work was so difficult and because of the pressures of competition for funding and other aspects of academic life (e.g. lack of time for study), the field fragmented, and as more people became involved the research community became more fragmented, with each group investigating only a small subset of the larger whole, and talking only to members of that group.
- Deep, narrowly focused, research on very specific problems is a requirement for progress, but if everybody does only that, the results will be bad.
  - People working on natural language without relating it to studies of perception, thinking, reasoning, and acting may miss out on important aspects of how natural languages work.
  - Likewise those who study only a small sub-problem in perception may miss out ways in which the mechanisms they study need to be modified to fit into a larger system.
  - The study of emotions also needs to be related to the total system.

The European Community's initiative in 'Cognitive Systems' (begun 2003) was an attempt to remedy this by requiring researchers to think about integrated multi-component systems. But there were not enough well-educated researchers.
One of the projects funded (including Birmingham) under that initiative is described here:
`http://www.cs.bham.ac.uk/research/projects/cosy/`

A UK grand challenge proposal to put all the pieces together again in a long term research programme is described here `http://www.cs.bham.ac.uk/research/cogaff/gc/`

Example demos

Some 'toy' examples of this design-based approach were shown during the talk.

They included

- The simulated 'harassed nursemaid' having to look after too many 'babies' in an environment presenting various opportunities and dangers
- Two simulated 'emotional' individuals trying to get to their 'targets' and becoming glum, surprised, neutral, or happy depending on what happened in their toy world: these have knowledge of their own states (unlike the nursemaid) and express the state both in a change of facial expression and a verbal report.
- A simulated sheepdog which fetches sheep and herds them into a pen (one at a time) in a world in which its plans can be blocked (e.g. because a tree is moved to block its path, or it or one of the sheep can be forcibly moved to a new location, requiring it to abandon its current plan and form a new one), and in which new opportunities can turn up unexpectedly (e.g. because a barrier that required a long detour suddenly acquires a gap, allowing the dog to use a short-cut). This dog has no anger or frustration when things go wrong, or joy when new opportunities suddenly appear: but it is able to detect new developments and react to them appropriately.

There are movies showing these programs online here

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

However, these are all toy systems: all they do is illustrate the design-based approach (for nursery school learners??).

Anyone who wishes to acquire and play with the software tools can fetch them from here

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

The tools require a linux or solaris system.

Linux running under vmware on a windows or mac pc may work.

Conceptual confusions: how to make progress

- We may be able to come up with clear, useful design-based concepts for describing what is happening in a certain class of complex information-processing systems, if we study the architecture, mechanisms and forms of representations used in that type of system, and work out the states and processes that can be generated when the components interact with each other and the environment.
- If the system is one that we had previously encountered and for which we already have a rich and useful pre-scientific vocabulary, then the new design-based concepts will not necessarily replace the old ones but may instead refine and extend them, e.g. making new sub-divisions and bringing out deep similarities between previously apparently different cases.
- This happened to our concepts of physical stuff (air, water, iron, copper, salt, carbon, etc.) as we learnt more about the underlying architecture of matter and the various ways in which the atoms and sub-atomic particles could combine and interact. So we now define water as $\mathbf{H_2O}$ and salt as $\mathbf{N_aC_l}$, rather than in terms of how they look, taste, feel, etc., and we know that there are different isotopes of carbon with different numbers of neutrons, e.g. $\mathbf{C_{12}}$, $\mathbf{C_{13}}$ and $\mathbf{C_{14}}$.
- As we increase our understanding of the architecture of mind (what the mechanisms are, how they are combined, how they interact) our concepts of mind (e.g. 'emotion', 'consciousness', 'learning', 'seeing', etc.) will also be refined and extended.

In the meantime, muddle and confusion reign.

Varieties of definitions of emotion

Part of the problem is that many of the words we use for describing human mental states and processes (including 'emotion', 'learning', 'intelligence', 'consciousness') are far too ill-defined to be useful in scientific theories.

Not even professional scientists are close to using agreed definitions of 'emotion'.

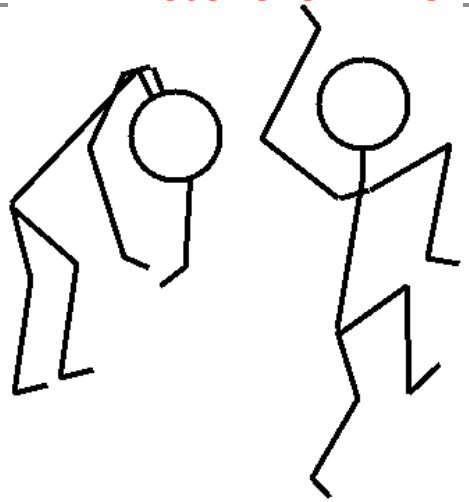
In the psychological literature, for instance, there are attempts to define emotions in terms of

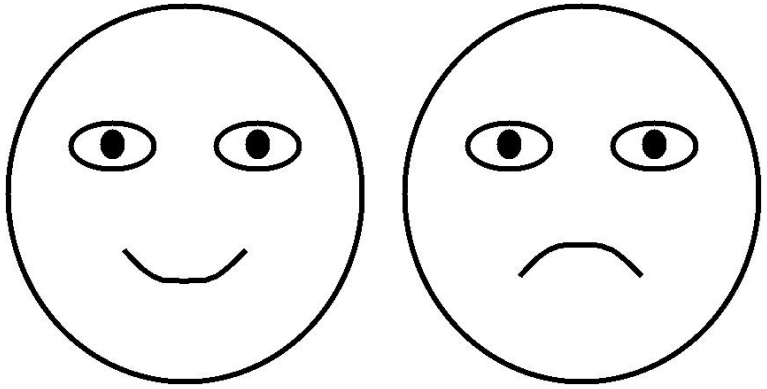
- social relations and interactions between people
- the kinds of states in the environment that produce them
- the kinds of behaviours they produce
- kinds of input-output relations (combining both the above)
- 'skin-level' and 'sub-skin-level' states and processes, e.g. whether hairs stand on end, galvanic skin responses, blood pressure, muscular tension, etc.
- the experience of the above bodily changes due to proprioceptive feedback mechanisms (the James/Lange definition, revived in a slightly new form by Damasio's theory of 'somatic markers')
- which bits of the brain produce them (e.g. amygdala, ...)
- 'how it feels' to have them
- how they affect other aspects of mental life
- Compare Sartre: to have an emotion is to see the world as "magical".

  .....etc....

All this conceptual confusion and definitional disagreement makes it unclear what question we are asking when we ask whether emotions are needed for intelligence.

What do we mean by "having an emotion"?





- Is it enough to produce certain behaviours that people interpret as emotional?
- Do actors actually have the states they portray so effectively — e.g. despondency, joy, jealousy, hatred, grief...? Not when such states include beliefs and intentions, as despondency, joy, jealousy, hatred, grief etc., often do.

- Behaviour is not enough to define any mental state, since
- In principle any behaviour, observed over any time period, can be produced by indefinitely many different mechanisms, using very different internal states and processes. Hence the Turing test is of no use here.
- We need to understand the variety of types of mental states better.
  Then we can define scientific concepts for classifying such states.

We need a "design-based", in this case "architecture-based" theory of emotions. (Contrast "dimensions-based" theories.)

METHODOLOGICAL POINT

The concept of emotion is but one of a large family of intricately related, but somewhat confused, everyday 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 the notion of 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 limited 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 for minds,
e.g. infants, toddlers, children, teenagers, professors ....
each providing its own family of concepts.

So we need many periodic tables
generating different sets of concepts.

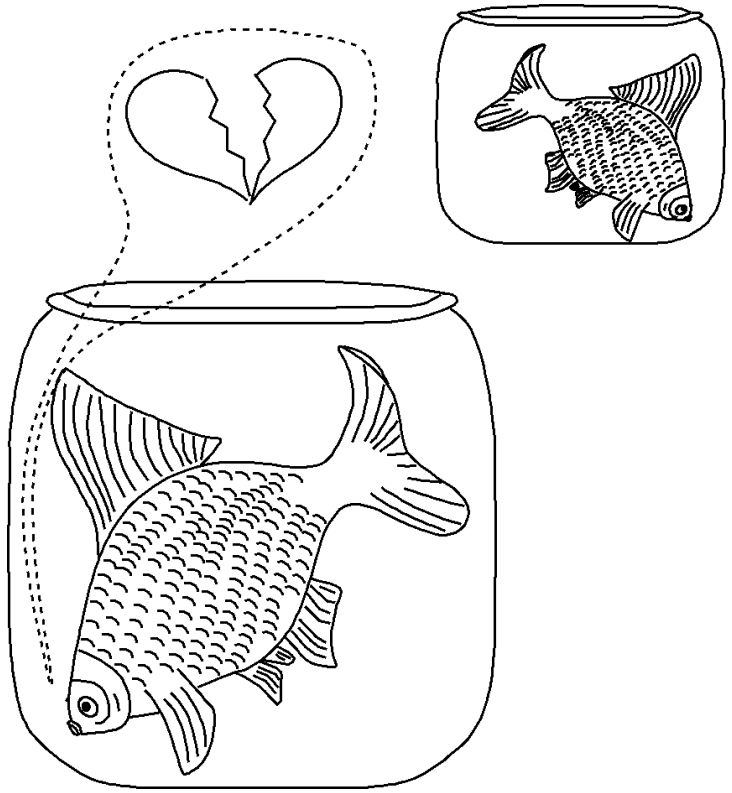
There may be some concepts applicable across architectures

What's wrong with the concepts?

- Everyday concept of 'emotion' mixes up motivations, attitudes, preferences, evaluations, moods, and other affective states and processes.
- There's not even agreement on what sorts of things can have emotions
  - A fly?
  - A woodlouse?
  - A fish?
  - An unborn human foetus?
  - An operating system?
  - A nuclear power plant warning system?
- **E.g. some people who argue that emotions are needed for intelligence are merely defending David Hume's truism that motivation is needed for action (though not in the case of tornadoes), and preferences are needed for selecting between options.**

  Does a tornado **select** a direction to move in?

  Does a paramoecium?



Towards a general framework

We need to talk about "information-using systems" — where "information" has the everyday sense (Jane Austen's sense[1]) not Shannon's technical sense. This notion is being used increasingly in biology.

What are information-using systems?

- They acquire, store, manipulate, transform, derive, apply information.
- The information must be expressed or encoded somehow, e.g. in simple or complex structures – possibly in virtual machines.

  (The use of *physical* symbol systems is often too restrictive.)
- These information-bearers may be within the system or in the environment.
- The information may be more or less explicit, or implicit.

A theory of meaning as we normally understand "meaning" in human communication and thinking should be seen as a special case within a general theory of information-using animals and machines.

These ideas are explained in more detail, including the notion of information processing in virtual machines here:

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#models`
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#wpe08`
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html`

[1] `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/austen-info.html`

Examples of types of process involving information

- Acquisition
- Filtering/selecting
- Transforming/interpreting/disambiguating
- Compressing/generalising/abstracting
- Proving, deriving (making inferences, but not only using propositions)
- Storage/Retrieval (many forms: exact, pattern-based, fuzzy)
- Training, adaptation (e.g. modifying weights, inducing rules)
- Constructing (e.g. descriptions of new situations or actions)
- Comparing and describing information (meta-information)
- Reorganising (e.g. formation of new ontologies)
- Testing/interrogating (is X in Y, is A above B, what is the P of Q?)
- Copying/replicating
- Syntactic manipulation of information-bearing structures
- Translating between forms, e.g. propositions, diagrams, weights
- Controlling/triggering/modulating behaviour (internal, external)
- Propagating (e.g. in a semantic net, or neural net)
- Transmitting/communicating
- .... (many more)

The differences involve: types of content, types of medium used, and the causal and functional relations between the processes and their precursors and successors in causal chains.

Control information vs factual information

A feature of ordinary language that can confuse discussions of information-processing is that we normally think of information as something that is true or false: e.g. information about when the train will arrive, whereas much information is control information which instead of being a potential answer to a question about what is the case is a potential answer to a question about what to do (or not do).

Gilbert Ryle (*The Concept of Mind* 1949) distinguished knowing that and knowing how, and we could add knowing what to do, or avoid, or refrain from, or...

Examples include:

- recipes and instruction manuals
- the ten commandments
- books on etiquette
- commands given by superiors to subordinates
- advice given by parents to children or helpers to friends
- learnt skills that enable us to do things, ...

Control information is more fundamental to intelligent action, or any kind of action, than factual information, since control information can generate action without factual information, whereas the converse is not true (as David Hume and others noted).

Having motives, having preferences, having values, having attitudes, having ideals, having dislikes, all involve control information – in the virtual machines constituting minds – but there's no reason to regard them all as 'emotions'.

The importance of virtual machines

During the 20th century computer scientists and software engineers came to realise this important truth:

In addition to physical machines, whose components and behaviours can be described using the language of the physical sciences, e.g. physics and chemistry, there are also virtual machines whose components and behaviour require a quite different vocabulary for their description, and whose laws of behaviour are not like physical laws.

For more on this see

`http://www.cs.bham.ac.uk/research/cogaff/talks/#wpe08`

Virtual machines have many advantages over physical machines for various kinds of task, e.g.

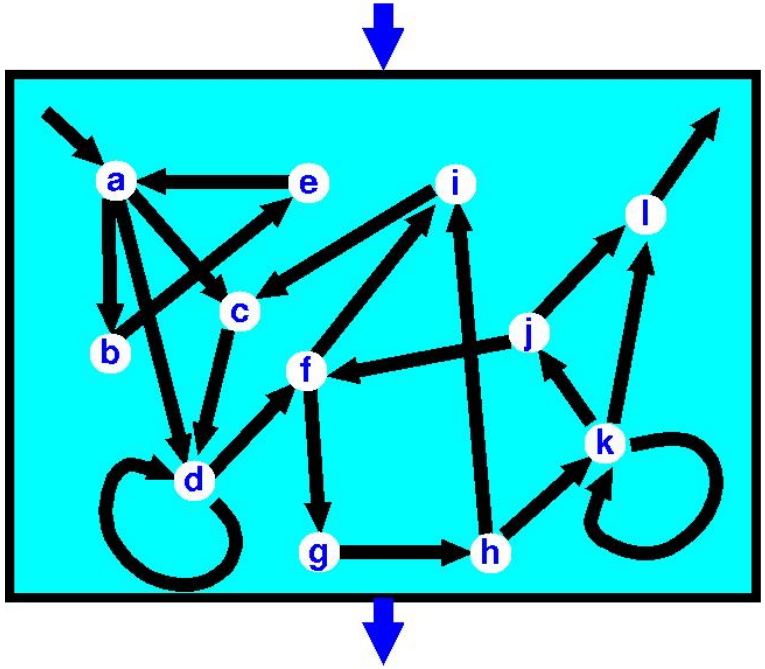
- They can change their structures without having to rebuild the underlying physical machinery
- They can switch between states containing different structures very quickly, far more quickly than physical structures can be reorganised

  This is needed for instance when what you see changes rapidly, or while you are having a rapid succession of complex thoughts, e.g. while reading a story or this text.
- Conflicts between inconsistent control processes can be resolved by deliberation and reasoning, instead of being restricted to numerical operations, e.g. averaging, or vector addition, as is the case with most physical forces pulling in different directions.

It is clear that evolution 'discovered' the benefits of virtual machines long before human scientists and engineers did!

Functionalism ?

Functionalism is one kind of attempt to understand the notion of virtual machine, in terms of states defined by a state-transition table.



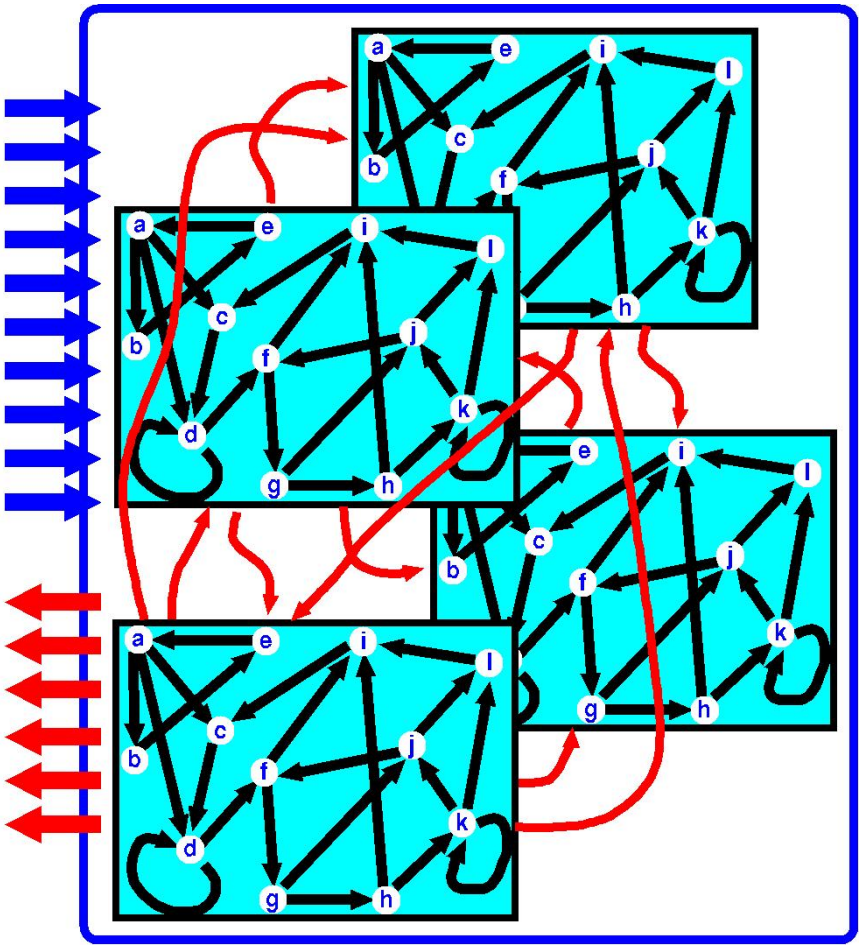
This is how many people think of functionalism: there's a total state which affects input/output contingencies, and each possible state can be defined by how inputs determine next state and outputs.

(E.g. see Ned Block's accounts of functionalism.)

HOWEVER THERE'S A RICHER, DEEPER NOTION OF FUNCTIONALISM

Another kind of Functionalism ?

Instead of a single (atomic) state which switches when some input is received, a virtual machine can include many sub-systems with their own states and state transitions going on concurrently, some of them providing inputs to others.



- The different states may change on different time scales: some change very rapidly others very slowly, if at all.
- They can vary in their granularity: some sub-systems may be able to be only in one of a few states, whereas others can switch between vast numbers of possible states (like a computer's virtual memory).
- Some may change continuously, others only in discrete steps.

Some sub-processes may be directly connected to sensors and effectors, whereas others have no direct connections to inputs and outputs and may only be affected very indirectly by sensors or affect motors only very indirectly (if at all!).

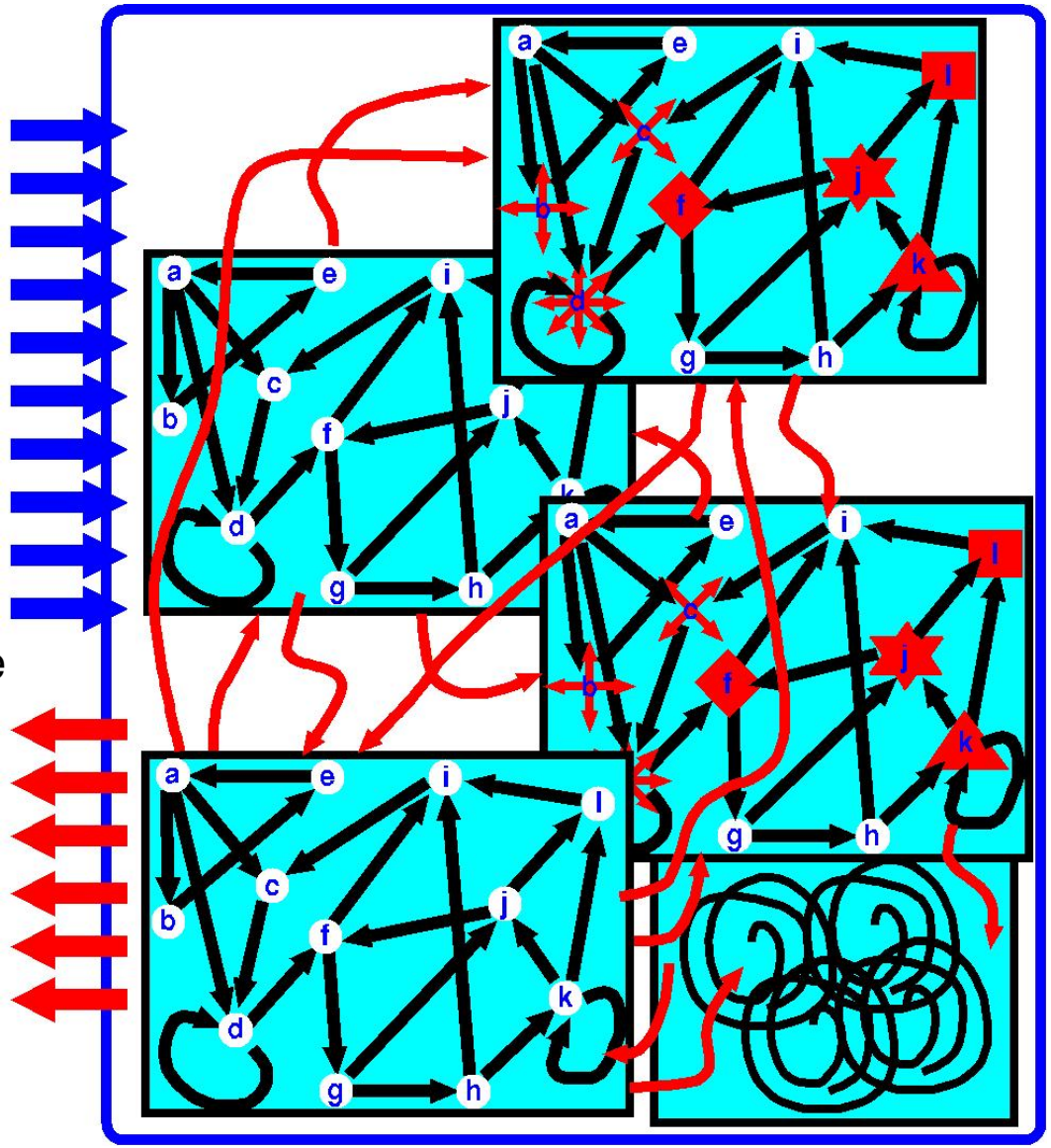
The previous picture is misleading

Because it suggests that the total state is made up of a fixed number of discretely varying sub-states:

We also need to allow systems that can grow structures whose complexity varies over time, as crudely indicated on the right,
e.g. trees, networks, algorithms, plans, thoughts, etc.

And systems that can change continuously, such as many physicists and control engineers have studied for many years, as crudely indicated bottom right
e.g. for controlling movements.

The label 'dynamical system' is trivially applicable to all these types of sub-system and to complex systems composed of them: but it explains nothing.



VMF: Virtual Machine Functionalism

We use "Virtual Machine Functionalism" (VMF) to refer to the more general notion of functionalism, in contrast with "Atomic State Functionalism" (ASF) which is generally concerned with finite state machines that have only one state at a time.

- VMF allows multiple concurrently active, interactive, sub-states changing on different time scales (some continuously) with varying complexity.
- VMF also allows that the Input/Output bandwidth of the system with multiple interacting internal states may be too low to reveal everything going on internally.
- There may still be real, causally efficacious, internal virtual machine events and processes that cannot be directly observed and whose effects may not even be indirectly manifested externally.

  Even opening up the system may not make it easy to observe the VM events and processes (decompiling can be too hard). See
  `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html`
  `http://www.cs.bham.ac.uk/research/cogaff/talks/#inf`
  `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#wpe08`
- VMF allows some processes to have the effect of providing control information for others, and for different processes to compete for control.
- Calling all control 'emotional' makes the label vacuous: but recognizing some special cases as emotional may be useful, e.g. some of the cases where one process actually or potentially disrupts, aborts, suspends, or otherwise interferes with another — e.g. when perceiving a predator suppresses foraging, or we are 'moved' by something.

CogAff: A schema for a variety of architectures.

'CogAff' is our label, not for an architecture (like 'H-Cogaff'), but for a way of specifying architectures in terms of which sorts of components they include and how they are connected: H-Cogaff is a special case of the schema.

Think of a grid of co-evolved types of sub-organisms, each contributing to the niches of the others, each performing different functions, using different mechanisms, etc.

We could add lots of arrows between boxes indicating possible routes for flow of information (including control signals) – in principle, mechanisms in any two boxes can be connected in either direction.

However, not all organisms will have all the kinds of components, or all possible connections.

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

E.g. insects are purely reactive, and perhaps also all reptiles and fish. A few species have deliberative capabilities in a simple form and perhaps even fewer have meta-management. Many kinds need "alarm" mechanisms.

For a survey of varieties of deliberative systems from 'proto-deliberative' to 'fully deliberative' see
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604`

Our own work in Birmingham
The architecture of a human mind

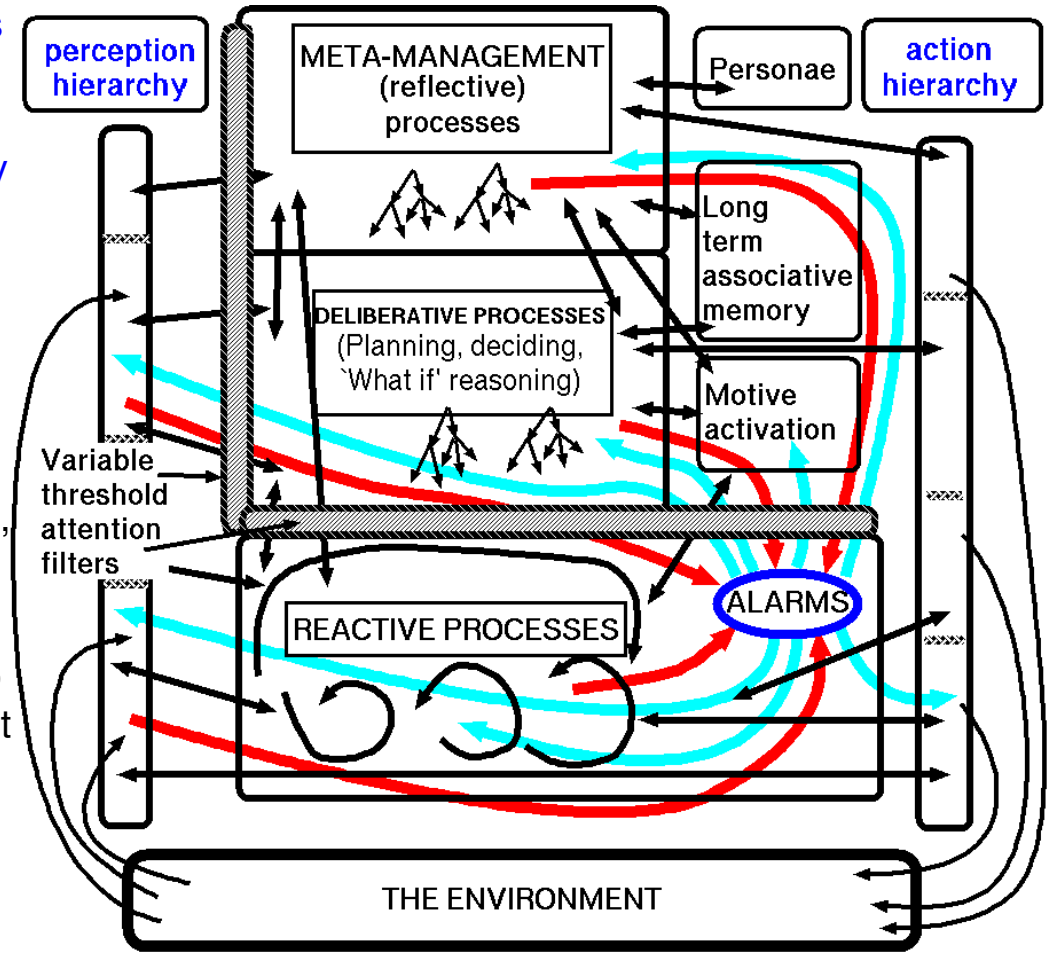
The Cognition and Affect project

(sketchy summary – see http://www.cs.bham.ac.uk/research/cogaff/#overview)

The H-Cogaff (Human Cogaff) architecture is a (conjectured) special case of the CogAff architecture schema, containing many different sorts of concurrently active, mutually interacting components.

It includes 'old' reactive components shared with many other animals (most species are purely reactive) 'newer' deliberative mechanisms (for considering non-existent possibilities) and relatively rare meta-management capabilities for inspecting, evaluating, and influencing internal information-processing.

Papers and presentations on the CogAff web site (URL above) give more information about the functional subdivisions in the (still very sketchy) H-Cogaff architecture, and suggest that many familiar kinds states (e.g. several varieties of emotions) could arise in such an architecture, in animals or robots. Long term grief is discussed in (Wright et al., 1996)).



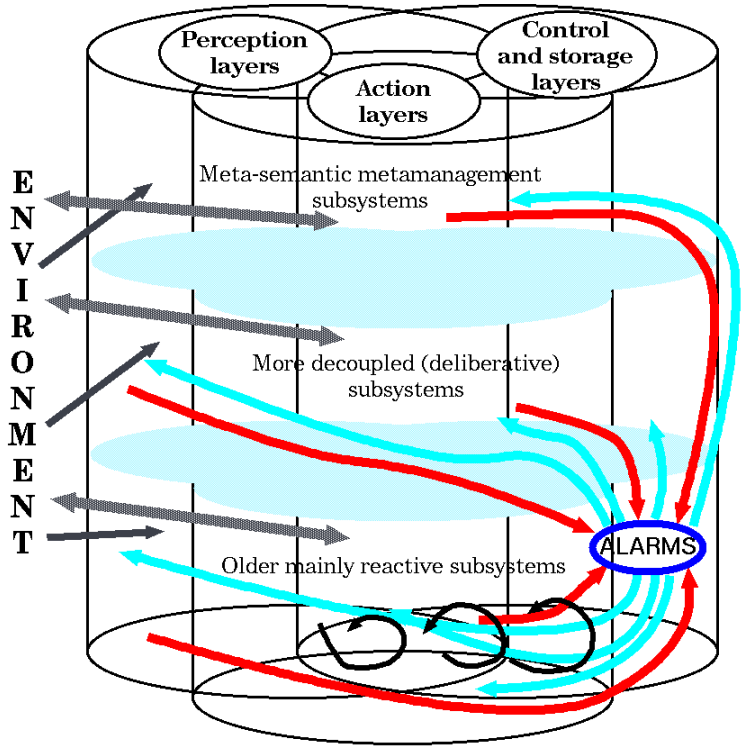
See other Cognition and Affect papers and talks for details

Overlapping architectural pillars

The previous diagrams suggest that sensory subsystems and motor subsystems are separate from each other with a “central” processing subsystem between them.

Many architectural theories make this sort of assumption – for instance researchers and robot designers who think of how brains work in terms of sensory-motor loops, e.g. (Powers, 1973)

It has been obvious for a long time that this is a serious error, since sensory and motor subsystems are closely connected, e.g. when eye movements are part of the process of visual perception and hand movements are part of haptic or proprioceptive perception.
See also (Gibson, 1966).



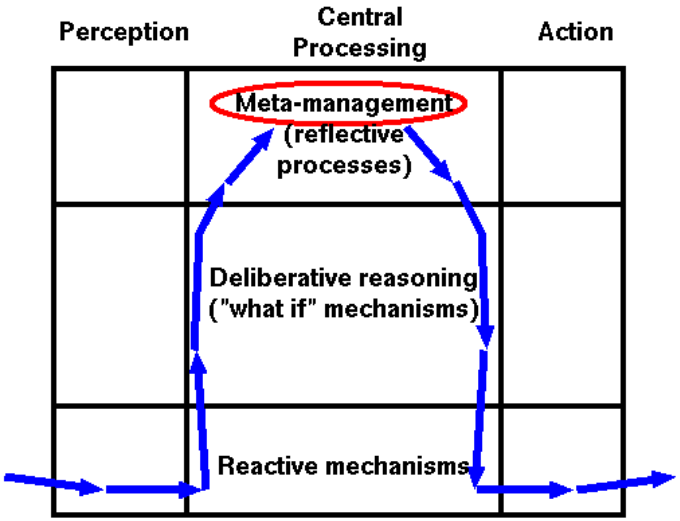
This diagram, produced with help from Dean Petters, is an attempt to remedy the flaw in the earlier diagrams, by indicating the overlaps between sensory and motor systems.

The notion of a collection of sensory-motor loops also leaves out many aspects of mental function that are disconnected from action, e.g. enjoying a musical performance without dancing, listening to a story, watching and thinking about what others are doing, constructing a poem or a mathematical proof with eyes shut and without performing any external actions.

This leads to the notion of a mind as involving many different dynamical systems performing a variety of tasks in parallel at different levels of abstraction along with a huge variety of mostly dormant but potentially active subsystems that can be made active either by external triggers, or by internal processes.

Another error: “peephole” perception and action

A different kind of error regards sensory and motor subsystems as “low level” input and output devices, dealing only with incoming physical stimuli and outgoing physical effects, as indicated in the diagram below, depicting what could be called an “Omega” architecture, because the information flow pattern resembles the Greek Ω.

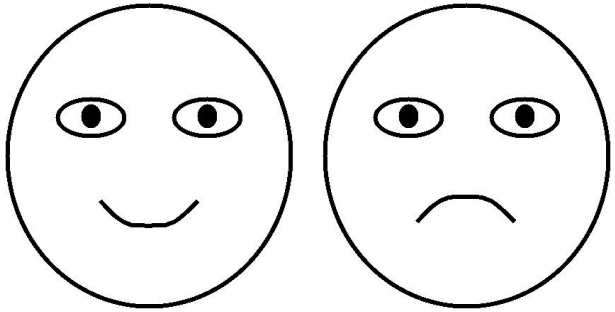


Such designs ignore the fact that biological evolution plus learning and development enable perception and motor control to become increasingly sophisticated, for example when vision includes parsing of text or grouping and interpretation of complex scene structures and processes (Johansson, 1973); and action includes sophisticated performances such as speaking while walking, painting a picture, or interacting socially using multiple simultaneous levels of control of movements.

An example of the way in which higher level processes can be part of the perceptual mechanism and produce results that are in registration with the input field is the picture of two faces.

Many people see the eyes in the two faces as different, even though they are geometrically identical.

The differences in the two mouths seem to produce emotionally different interpretations of the two faces, but that emotionality can influence how the other parts of the face (e.g. eyes) are perceived. (“Perception is controlled hallucination.” (Clowes, 1971))



This is related to the ideas presented above about “virtual machine functionalism”. Also: emotions can be in the eye of the beholder.

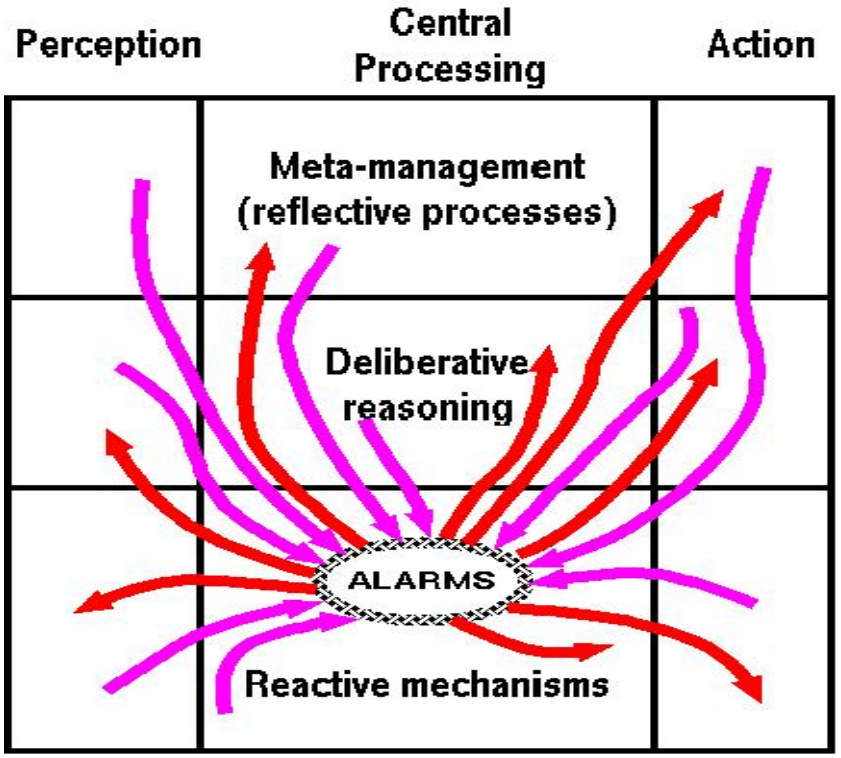
As processing grows more sophisticated, so it can become slower, to the point of danger

REMEDY: FAST, POWERFUL, "GLOBAL ALARM SYSTEMS"

Resource-limited 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 state can be based on such an alarm system, depending on what else is in the architecture.

Don't confuse the alarms (and emotions they produce) with the evaluations that trigger them, or the motives, preferences, policies, values, attitudes that have different sorts of functional roles – different sorts of control functions (including conditional control in many cases).

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 or more specialised sub-systems, somehow interrupt or suppress or change the behaviour of others, producing some alteration in (relatively) global (internal or external) behaviour of the system — which could be in a virtual machine.

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.

Some such 'emotional' states are useful, others not: they are not required for all kinds of intelligence — only in a subset of cases where the system is too slow or too uninformed to decide intelligently what to do — they can often be disastrous!

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.
  The two cases differ in what has been labelled "direction of fit".

Without affect there is no reason to do anything.

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

So I am NOT arguing that knowledge and powers of reasoning suffice for behaving intelligently: in particular, without motivation nothing will be done (except in purely reactive systems).

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

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

This is just the beginning

- I have tried to give some of the flavour of the kind of thinking involved in the design-based approach to thinking about minds of humans, other animals or machines.
- When we start investigating what could happen in an architecture as rich as H-Cogaff (which is still much simpler than a normal adult human architecture) we see that many more kinds of states and processes are possible than we have convenient labels for.
- So we can start classifying in a more precise way than ever before various classes of states and processes.
- We'll see that a subset of the things we call being in an emotional state (e.g. being startled, frightened of a cliff-edge, joyful at recognition of a loved one) may involve operations of something like the 'alarm' mechanism, though not all cases will be alike.
- Some of the long-term cognitively rich emotions including grief or jealousy may not depend on alarm mechanisms, likewise many attitudes often confused with emotions, e.g. dedication to one's job, love of one's family or country.

The periodic table of human mental states still has far to grow.

The ideas sketched here are, in part, a development of ideas that can be found in

H. A. Simon, (1967) Motivational and emotional controls of cognition, Reprinted in *Models of Thought*, Yale University Press, 29–38, 1979

See also: The Meta-Morphogenesis project:
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html`

For more on this approach
SEE THE COGNITION AND AFFECT PROJECT

OVERVIEW, INCLUDING PAPERS & DISCUSSION NOTES:

Overview of the Cognition and Affect Project, Birmingham.

`http://www.cs.bham.ac.uk/research/projects/cogaff/#overview`
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/emotions-questions.html`
`http://www.cs.bham.ac.uk/research/projects/cogaff/`
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/`

(References to other work can be found in papers in these directories)

TOOLS: (Poplog and the SIM_AGENT toolkit)

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

`http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html`

DEMO-MOVIES:

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

SLIDES FOR TALKS:

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/`
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#presentations`

(Including several on emotions)

ALSO RECOMMENDED:

(Wierzbicka, 1992) `http://csjarchive.cogsci.rpi.edu/1992v16/i04/p0539p0581/MAIN.PDF`
An important cross-cultural survey of labels used.

(Barrett, 2006) `http://www2.bc.edu/~barretli/pubs/2006/Barrett2006kinds.pdf`

And many more!

Some more references

NOTE:

1. Magda Arnold's edited collection gives a useful overview of emotion theories before 1968.
2. The Stanford Encyclopedia of Philosophy entry (de Sousa) also includes a useful bibliography.
3. See the BICA (Biologically Inspired Cognitive Architectures) Society web site:

   Biologically Inspired Cognitive Architectures Society
   `http://bicasociety.org/`
   Toward a Comparative Repository of Cognitive Architectures, Models, Tasks and Data
   `http://bicasociety.org/cogarch/`
   "the BICA Society shall bring together researchers from disjointed fields and communities who devote their efforts to solving the same challenge, despite that they may "speak different languages". This will be achieved by promoting and facilitating the transdisciplinary study of cognitive architectures, and in the long-term perspective - creating one unifying widespread framework for the human-level cognitive architectures and their implementations."

   I suspect many of the architectures proposed by AI theorists, and also some proposed by neuroscientists, can fit into the CogAff framework, but they all use different diagrammatic and notational conventions, making comparisons difficult.

References

Arbib, M., & Fellous, J.-M. (Eds.). (2005). *Who Needs Emotions?: The Brain Meets the Robot*. New York: Oxford University Press.

Arbib, M. A. (2002). From Rana Computatrix to Homo Loquens: A computational neuroethology of language evolution. In R. I. Damper et al. (Eds.), *WGW'02 Biologically-inspired robotics: The legacy of W. Grey Walter* (pp. 12–31). Bristol: Hewlett Packard Research Labs.

Arnold, M. B. (Ed.). (1968). *The nature of emotion*. Harmondsworth, England: Penguin Books.

Barrett, L. F. (2006). Emotions as natural kinds? *Perspectives on Psychological Science*, *1*(1), 28–58. Available from `http://affective-science.org/pubs/2006/Barrett2006kinds.pdf`

Beaudoin, L. (1994). *Goal processing in autonomous agents*. Unpublished doctoral dissertation, School of Computer Science, The University of Birmingham, Birmingham, UK. Available from `http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#38`

Beaudoin, L., & Sloman, A. (1993). A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, & A. Ramsay (Eds.), *Prospects for artificial intelligence* (pp. 229–238). Amsterdam: IOS Press. Available from `http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#16`

Boden, M. A. (1972). *Purposive Explanation In Psychology*. Cambridge, MA: Harvard University Press.

Chappell, J., & Sloman, A. (2007). Natural and artificial meta-configured altricial information-processing systems. *International Journal of Unconventional Computing*, *3*(3), 211–239. Available from `http://www.cs.bham.ac.uk/research/projects/cogaff/07.html#717`

Clowes, M. (1971). On seeing things. *Artificial Intelligence*, *2*(1), 79–116. Available from `http://dx.doi.org/10.1016/0004-3702(71)90005-1`

Damasio, A. (1994). *Descartes' error, emotion reason and the human brain*. New York: Grosset/Putnam Books.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London: Harper Collins. ((Reprinted 1998))
Donald, M. (2002). *A mind so rare: The evolution of human consciousness*. W W Norton & Company Incorporated. Available from http://books.google.co.uk/books?id=Zx-MG6kpf-cC
Frijda, N. H. (1986). *The emotions*. Cambridge: Cambridge University Press.
Ganti, T. (2003). *The Principles of Life* (E. Szathmáry & J. Griesemer, Eds.). New York: OUP. (Translation of the 1971 Hungarian edition, with notes)
Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.
Goleman, D. (1996). *Emotional intelligence: Why it can matter more than iq*. London: Bloomsbury Publishing.
Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, *14*, 201–211.
LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.
Minsky, M., Singh, P., & Sloman, A. (2004). The St. Thomas common sense symposium: designing architectures for human-level intelligence. *AI Magazine*, *25*(2), 113–124. (http://web.media.mit.edu/~push/StThomas-AIMag.pdf)
Minsky, M. L. (1987). *The society of mind*. London: William Heinemann Ltd.
Minsky, M. L. (2006). *The Emotion Machine*. New York: Pantheon.
Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of the emotions*. New York: Cambridge University Press.
Picard, R. (1997). *Affective computing*. Cambridge, MA; London, England: MIT Press.
Powers, W. T. (1973). *Behavior, the Control of Perception*. New York: Aldine de Gruyter.
Sartre, J. (1962). *Sketch for a Theory of the Emotions*. London: Methuen,. (Publ. 1939 in French)
Scheutz, M., & Sloman, A. (2001). Affect and agent control: Experiments with simple affective states. In *et al*. Ning Zhong (Ed.), *Intelligent Agent Technology: Research and Development* (pp. 200–209). New Jersey: World Scientific Publisher.
Simon, H. A. (1967). Motivational and emotional controls of cognition. In H. A. Simon (Ed.), *reprinted in models of thought* (pp. 29–38). Newhaven, CT: Yale University Press.
Sloman, A. (1978). *The computer revolution in philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press). Available from http://www.cs.bham.ac.uk/research/cogaff/crp
Sloman, A. (1982). Towards a grammar of emotions. *New Universities Quarterly*, *36*(3), 230–238. Available from http://www.cs.bham.ac.uk/research/cogaff/81-95.html#emot-gram
Sloman, A. (1987). Motives mechanisms and emotions. *Cognition and Emotion*, *1*(3), 217–234. (Reprinted in M.A. Boden (ed), *The Philosophy of Artificial Intelligence*, 'Oxford Readings in Philosophy' Series, Oxford University Press, 231–247, 1990)
Sloman, A. (1992). Prolegomena to a theory of communication and affect. In A. Ortony, J. Slack, & O. Stock (Eds.), *Communication from an artificial intelligence perspective: Theoretical and applied issues* (pp. 229–260). Heidelberg, Germany: Springer. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#10
Sloman, A. (1994). Computational modelling of motive-management processes. In N.Frijda (Ed.), *Proceedings of the conference of the international society for research in emotions* (pp. 344–348). Cambridge: ISRE Publications.
Sloman, A. (1999). Beyond shallow models of emotion. In E. Andre (Ed.), *Behaviour planning for life-like avatars* (pp. 35–42). Sitges, Spain. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/00-02.html#74 (Proceedings I3 Spring Days Workshop March 9th–10th 1999)
Sloman, A. (2001a). Beyond shallow models of emotion. *Cognitive Processing: International Quarterly of Cognitive Science*, *2*(1), 177–198.
Sloman, A. (2001b, March). Varieties of Affect and the CogAff Architecture Schema. In C. Johnson (Ed.), *Proceedings Symposium on Emotion, Cognition, and Affective Computing AISB'01 Convention* (pp. 39–48). York.
Sloman, A. (2002). How many separately evolved emotional beasties live within us? In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in Humans and Artifacts* (pp. 35–114). Cambridge, MA: MIT Press.
Sloman, A. (2004, March 2004). What are emotion theories about? In E. Hudlicka & L. C. namero (Eds.), *Proceedings Spring Symposium on Architectures*

for Modeling Emotion: Cross-Disciplinary Foundations (pp. 128–134). Menlo Park, CA: AAAI. (http://www.cs.bham.ac.uk/research/projects/cogaff/04.html#200403)

Sloman, A. (2006, May). *Requirements for a Fully Deliberative Architecture (Or component of an architecture)* (Research Note No. COSY-DP-0604). Birmingham, UK: School of Computer Science, University of Birmingham. Available from http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604

Sloman, A. (2009). Architecture-Based Motivation vs Reward-Based Motivation. *Newsletter on Philosophy and Computers*, *09*(1), 10–13. Available from http://www.apaonline.org/publications/newsletters/v09n1_Computers_06.aspx

Sloman, A., Chrisley, R., & Scheutz, M. (2005). The architectural basis of affective states and processes. In M. Arbib & J.-M. Fellous (Eds.), *Who Needs Emotions?: The Brain Meets the Robot* (pp. 203–244). New York: Oxford University Press. (http://www.cs.bham.ac.uk/research/cogaff/03.html#200305)

Sloman, A., & Croucher, M. (1981). Why robots will have emotions. In *Proc 7th int. joint conference on AI* (pp. 197–202). Vancouver: IJCAI. Available from http://www.cs.bham.ac.uk/research/cogaff/81-95.html#36

Sousa, R. de. (2013). Emotion. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2013 ed.). Available from http://plato.stanford.edu/archives/spr2013/entries/emotion/

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.*, *42*(2), 230–265. Available from http://www.thocp.net/biographies/papers/turing_oncomputablenumbers_1936.pdf

Turing, A. M. (1952). The Chemical Basis Of Morphogenesis. *Phil. Trans. R. Soc. London B 237*, *237*, 37–72.

Turner, T., & Ortony, A. (1992). Basic Emotions: Can Conflicting Criteria Converge? *Psychological Review*, *99*, 566–571. (3)

Ugur, E. (2010). *A Developmental Framework for Learning Affordances*. Unpublished doctoral dissertation, The Graduate School of Natural and Applied Sciences, Middle East Technical University, Ankara, Turkey. Available from http://www.cns.atr.jp/~emre/papers/PhDThesis.pdf

Wierzbicka, A. (1992). Defining Emotion Concepts. *Cognitive Science*, *16*, 539–581.

Wright, I. (1977). *Emotional agents*. Unpublished doctoral dissertation, School of Computer Science, The University of Birmingham. (http://www.cs.bham.ac.uk/research/cogaff/)

Wright, I., Sloman, A., & Beaudoin, L. (1996). Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, *3*(2), 101–126. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#2