Slides presented at
Goldsmiths University, London,
19 Jan 2005

A RESEARCH GRAND CHALLENGE

Architectures for Human-like Machines

Towards INTEGRATED understanding of

- High Level Mental Processes
- Brain Mechanisms and Functions.

Aaron Sloman

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

Based on the documents here:
http://www.cs.bham.ac.uk/research/cogaff/gc/
and the presentations here:
http://www.cs.bham.ac.uk/research/cogaff/talks/

THANKS

I am very grateful to
the developers of Linux
and other free, open-source,
platform-independent, software systems.

LaTex was used to produce these slides.

Diagrams are created using tgif, freely available from
http://bourbon.cs.umd.edu:8001/tgif/

Demos are built on Poplog

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

Summary of talk: 1

- Current ways of studying (animal, human and robot) minds are
  - too fragmented
  - too riddled by turf wars
  - too much influenced by prejudice (what people would like to be true)
  - based on inadequate notions of science and explanation
  - based on too little data in forms that are too restricted
- Examples:
  - bad theories about emotions
  - confused concepts treated as well understood
  - theories/models/explanations that don't 'scale out' (fit into a larger context)
- We can remedy this by working out the implications of these facts:
  - minds DO things: they are constantly active machines
  - there is not just one kind of mind: very many exist in nature, even among humans: young, old, normal, damaged, ancient, modern (industrialised)
  - all organisms are information processors
  - evolution is far ahead of our understanding
  - all complex designs involve complex trade-offs
  - new evolutionary designs do not simply throw away old solutions, but build on them:

    humans share much with much older species

Summary of talk: 2

- This project requires contributions from many disciplines – and they will be changed by contributing to the project.

- It can also build on common sense

  A new science of mind need not throw away the rough-hewn concepts of ordinary language, and the vast amount of folk-knowledge we use every day (mostly unconsciously, much embedded in our use of language), but can use them as stepping stones to a richer, deeper, collection of ways of thinking about what sorts of machines we are, and might be.

  Compare the way physics as deep explanatory science builds on and explains ‘folk physics’ instead of throwing it all away.

- A major contribution from Computer Science, AI and Software engineering is new understanding of levels of abstraction:

  layers of virtual machines running on physical machines.
  (But our understanding is still rudimentary).

- Understanding ourselves is an exercise in designing working systems – not just a matter of collecting, correlating, summarising, organising, observations

Some old ways to study minds

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

- **Reading plays, novels, poems** will teach you much about how people who see, act, 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 about how mental phenomena, including cognitive capabilities vary among different animals.
- **Studying psychology** will add much extra detail concerning what can be triggered or measured in laboratories, and what correlates with what.
- **Studying developmental psychology** can teach you how the states and processes in infants differ from those in older children and adults.
- **Neuroscience** teaches us about physiological brain mechanisms that support and modulate mental states and processes, and are modulated by them.
- **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 minds are and how mental states and processes differ from one another and from physical states and processes.

Another, less familiar, way complements and enriches 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 (using various sensory systems),
- learning (concepts, forms of representation, facts, generalisations, skills, ...)
- acquiring new motives, values, standards, preferences, ideals ...)
- enjoying some activities and states and disliking others,
- selecting between conflicting motives,
- planning, executing plans, planning how to plan, ...
- reacting to dangers and opportunities,
- communicating in various ways (including collaborating, competing and fighting)
- 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 (and other animals) 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 recent initiative in 'Cognitive Systems' is an attempt to remedy this by requiring researchers to think about integrated multi-component systems.
One of the projects to be 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/

The need for integration

Over the last half-century there has been much fragmentation, within each of: AI, psychology, neuroscience — most researchers focus only on a limited sub-field, e.g.

- vision (usually low-level vision nowadays)
- language (text, speech, sign-language)
- learning (many different kinds)
- problem solving
- planning
- mathematical reasoning
- motor control
- emotions

etc....

The need to 'scale out' is more important than the need to 'scale up'

There is no guarantee that a technique, or form of representation, or algorithm, etc. that works for an isolated task will also work when that task has to be integrated with many other kinds of functionality in an integrated system. This is true of AI techniques that 'scale up' very well within a particular application domain, e.g. path planning. E.g. they may not 'scale out' to support anytime planning or reasoning about planning, or cooperative planning.

Human abilities generally do not scale up: we are defeated by combinatorics.
But they scale out and interact fruitfully: e.g. what you see can help you understand words you hear and vice versa. (McGurk effect)

Putting the pieces back together:

- We need to understand and model brains/minds as integrated working systems functioning at different levels of abstraction, including
  - Physiological properties of brain mechanisms (how many different sub-types are there?)
  - Neural information processing functions
  - ‘Higher level’ cognitive and affective functions of many sorts, implemented in older mechanisms.
  - Behaviours of complete agents (including social behaviours).
- This requires us to understand how the different levels, and the different components at each level, combine to form an integrated functioning system
  - some levels implementing others,
  - some sub-systems cooperating with or competing with others
- We need to understand principles of operation at different levels rather than always merely trying to mimic low level biologicall details.

  Compare: the understanding of software engineers and digital electronic engineers, or physicists.
- That includes finding good characterisations of requirements for architectures, mechanisms, formalisms, at all levels ...

Using factual material

- One problem is identifying what needs explaining.

  Too often people observe only what their theories deem relevant, or collect only information that their statistical tools can process.

- A scenario-based approach can help to overcome that limitation

  by collecting and analysing very many real scenarios, organised according to their similarities and differences and ordered by complexity
  e.g. (of mechanisms, of information, of architectures, of representations needed).

Examples: collect and study videos of animals and children:

- Betty, the new caledonian crow, surprised researchers at the Oxford University Zoology department when she displayed an ability to make a hook out of a straight piece of wire, in order to fish a bucket containing food out of a tube: (http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm)

- An 18 month old child attempts to join two parts of a toy train by bringing two rings together instead of a ring and a hook, and showing frustration and puzzlement at his failure. (http://www.cs.bham.ac.uk/˜axs/fig/josh34_0096.mpg)
  A few weeks later he was able to solve the problem: what had changed?

- If time: video of the child playing with trains on the floor about a year later.

Supplement observed scenarios with a large collection of analytical scenarios: compare Piaget

THE GRAND CHALLENGE PROBLEM

How can we understand brains and minds of humans and other animals well enough to build convincing functional robot models?

Premises:

- **Understanding natural intelligence involves investigation at different levels of abstraction**
  - **Brain:**
    The physical machine, with physical, chemical, physiological and functional levels performing many different types of tasks in parallel including information-processing tasks and others (e.g. supplying energy).
  - **Mind:**
    The “virtual machine” (or collection of interacting virtual machines) performing many different types of information-processing tasks in parallel – at different levels of abstraction.
- **This is an enormously difficult long-term task, requiring cooperation between many disciplines, e.g.**
  - Neuroscience
  - Psychology
  - Computer science
  - Software engineering
  - AI (including Robotics)
  - Linguistics
  - Social sciences
  - Biology
  - Ethology
  - Anthropology
  - Philosophy
  - Mathematics / Logic

Two-way scientific information flow

We need a far better understanding of how natural intelligence works, at different levels of abstraction, if we are to build more intelligent (e.g. robust, autonomous, adaptive) artificial information-processing systems.

In particular, building working human-like robots requires us to develop architectures combining many types of functionality.

But in order to understand examples of natural intelligence we need to understand how to design systems with similar capabilities.

Eliminate ‘turf-wars’ between disciplines

Some people argue that explaining how humans and other animals work is a problem for biology, neuroscience and psychology, not computer science, e.g. because brains are not computers

But all organisms, including humans, are information-processing systems and there is no other discipline that has tools, techniques and theories for modelling and explaining a wide range of information- processing capabilities, especially in virtual machines.

So we can’t just leave this to other disciplines,
e.g. neuroscience, psychology —
but we must learn from and cooperate with them.

Science or Engineering?

This is primarily a scientific challenge, not an applications challenge aimed at producing some useful new machines.

But the research has two aspects, theoretical and practical, which inform each other.

POTENTIALLY THERE ARE MANY APPLICATIONS – BUT THEY ARE NOT THE MAIN MOTIVATION.

The engineering goal of getting a machine to play chess as well as the best human players has been achieved, but not the scientific goal of clarifying requirements and designs for a machine that understands what it is doing when it plays chess, and can describe its strategy, explain things to a weaker player, etc.

Impressive robots made by Honda and Sony

THE STATE OF THE ART IN 2002

http://www.aibo.com/

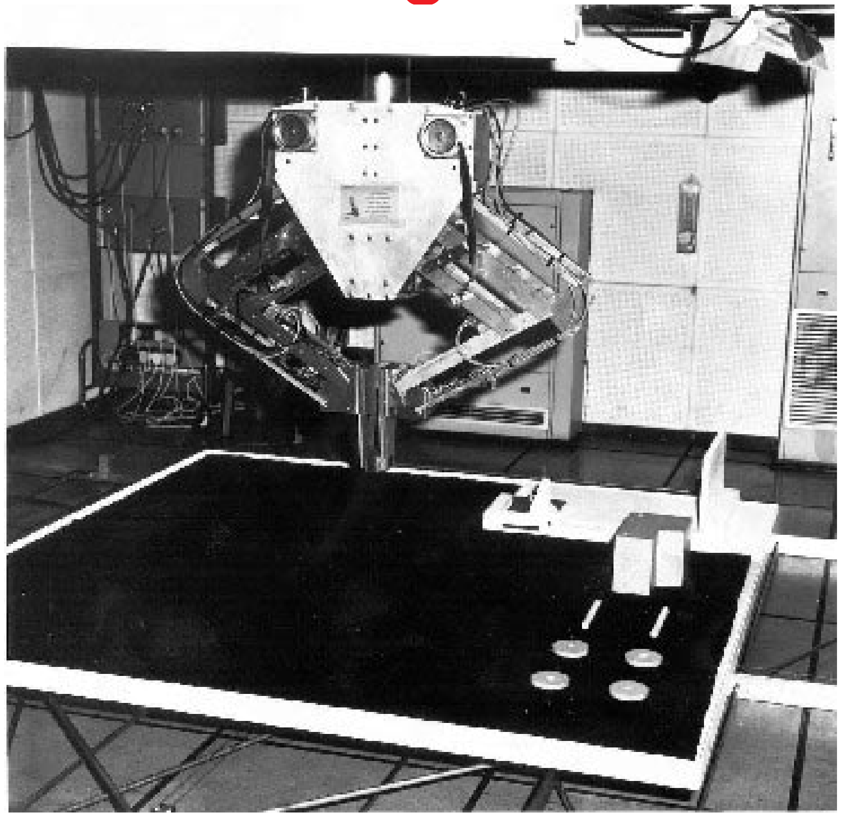
http://world.honda.com/news/2002/c021205.html

In both cases the engineering is very impressive. But present day robots look incompetent if given a task that is even slightly different from what they have been programmed to do – unlike a child or crow or squirrel.

Mostly they have purely reactive behaviours, lacking the deliberative ability to wonder 'what would happen if...'.

They also have very little or no self-knowledge or self-understanding, e.g. about their limitations, or why they do things as they do.

Compare Freddy the 1973 Edinburgh Robot



Some people might say that apart from wondrous advances in mechanical and electronic engineering there has been little increase in sophistication since the time of Freddy, the ‘Scottish’ Robot, built in Edinburgh around 1972-3.

Freddy II could assemble a toy car from the components (body, two axles, two wheels) shown. They did not need to be laid out neatly as in the picture.

However, Freddy had many limitations arising out of the technology of the time.

E.g. Freddy could not simultaneously see and act: partly because visual processing was extremely slow.

Imagine using a computer with 128Kbytes RAM for a robot now.

There is more information on Freddy here

http://www.ipab.informatics.ed.ac.uk/IAS.html

http://www-robotics.cs.umass.edu/ pop/VAP.html

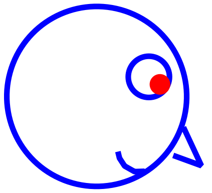
In order to understand the limitations of robots built so far, we need to understand much better exactly what animals do: we have to look at animals with the eyes of (software) engineers.

The time is ripe for a major new initiative

Many agree that because of recent advances in science and technology, and expected new advances in the next 20 years, the time is ripe for a new initiative.

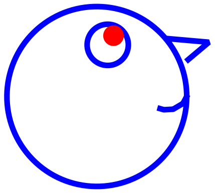
But there are different views as to how to proceed

- **Some prefer implementation-neutral “top-down” research strategies, attempting to explain known kinds of human and animal competence using any low-level mechanisms that work, whether biologically plausible or not.**



Top down

- **Some prefer to work “bottom” up from brain mechanisms (e.g. chemical and neural) along with physical properties of bodies.**



Bottom up

- **Solution:**
  - **a combined bottom-up, top-down and middle-out strategy,**
  - **using requirements at different levels of abstraction in combination, to constrain the search for good designs.**

Part of the scientific goal is to understand the requirements for integrated, multi-functional, human-like systems.
Many earlier projects mistakenly assumed the requirements were clear.

Architectural challenges

One requirement for progress is specification of a virtual machine architecture that can combine many known kinds of human capabilities, including

- evolutionarily very old reactive mechanisms, many shared with other species;
- newer deliberative mechanisms capable of representing and reasoning about what is not perceived, or might occur, or might have occurred;
- biologically rare reflective, meta-management mechanisms.

NB

- These are not mutually exclusive categories
- There are many intermediate cases – e.g. proto-deliberative reactive systems (contention scheduling)
- Fully deliberative systems are very rare – and biologically expensive
- Meta-management may use a mixture of the other kinds
- No type is dominant over others: all run in parallel and sometimes battle for supremacy.

Papers and presentations in the Cognition and Affect project provide more detailed analyses of these architectural features, illustrated on the next two slides. See

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

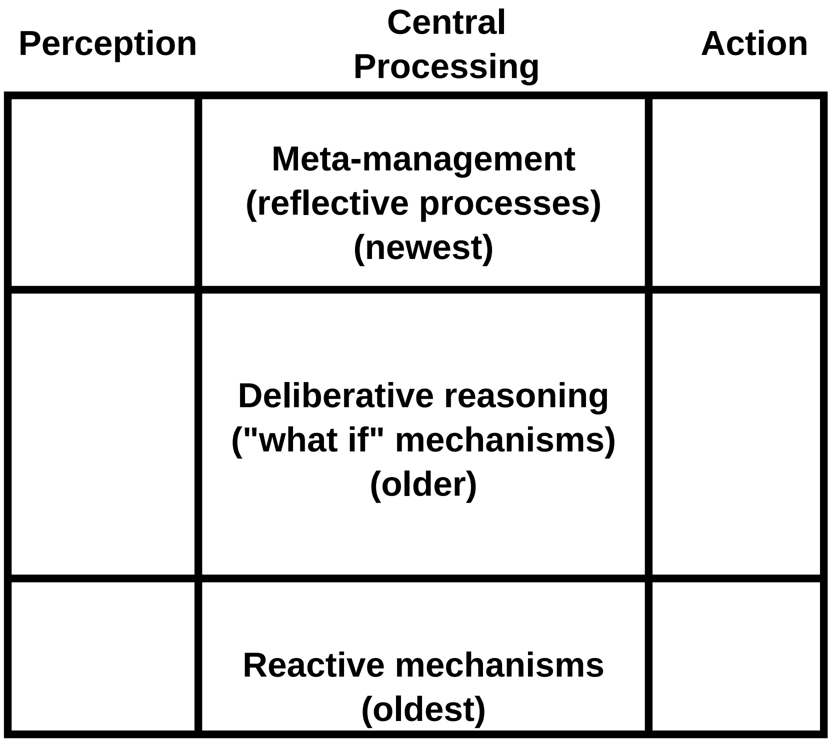
CogAff: A schema for a variety of architectures.

‘CogAff’ is our label, not for an architecture (like ‘H-Cogaff’, below), 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.

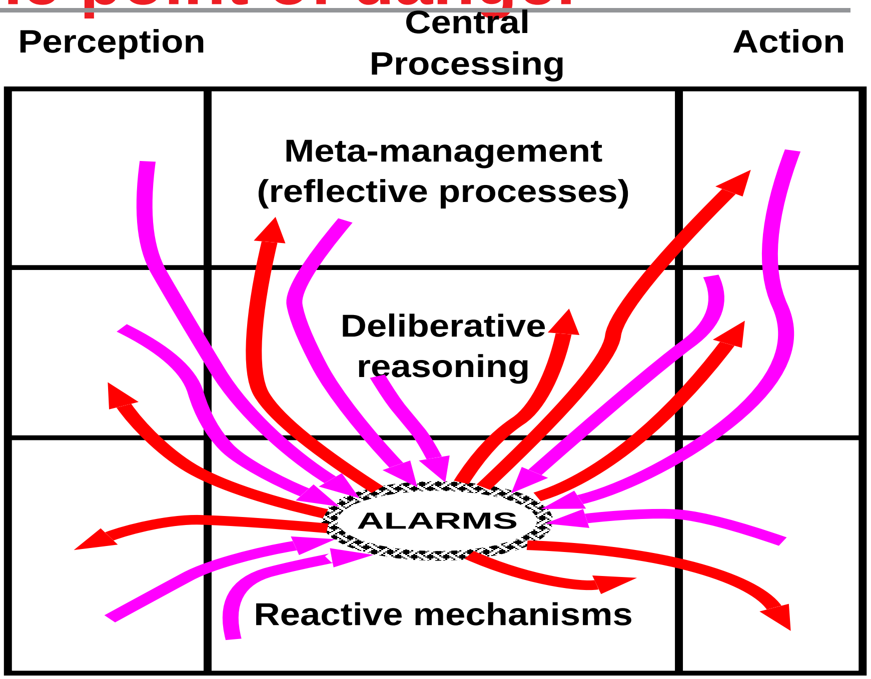
As processing grows more sophisticated, so it can become slower, to the point of danger

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, ...???)
See Cogaff papers and talks

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



Perception

Central Processing

Action

Meta-management (reflective processes)

Deliberative reasoning

ALARMS

Reactive mechanisms

Many different kinds of emotional states 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).

The ‘five Fs’

What we have called ‘alarm mechanisms’ may trigger behaviours often referred to as ‘the four Fs’, though there are at least five:

- fleeing
- fighting
- feeding
- freezing
- reproducing

(I added ‘freezing’ – often the best response to danger or uncertainty.)

In humans they can trigger far more complex and subtle processes including deliberative and meta-management processes (e.g. reasoning anxiously about whether it would be wise to continue pursuing one’s current goal).

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 (not original – e.g. see Herb Simon on emotions):

in all cases there are at least two sub-systems at work in the organism, and one or more specialised sub-systems can 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 ‘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!

This is not a semantic battle

- I am not arguing about what ‘emotion’ really means, or about what emotions really are.
- I am pointing to the need in some designs, in some environments, to have one or more types of ‘alarm’ mechanisms with the properties described.
- I claim that many familiar phenomena in humans and other animals, including some things we label ‘emotional’ could occur in systems with such a design.
- There are many phenomena that would not be explained by such a mechanism: e.g. enjoying solving a mathematical problem.

NOTE

In a famous paper on Truth the philosopher J.L.Austin once suggested that philosophers would make more progress if instead of arguing about what truth is they discussed what ’true’ means (adding “in vino veritas but in sober symposium verum”.

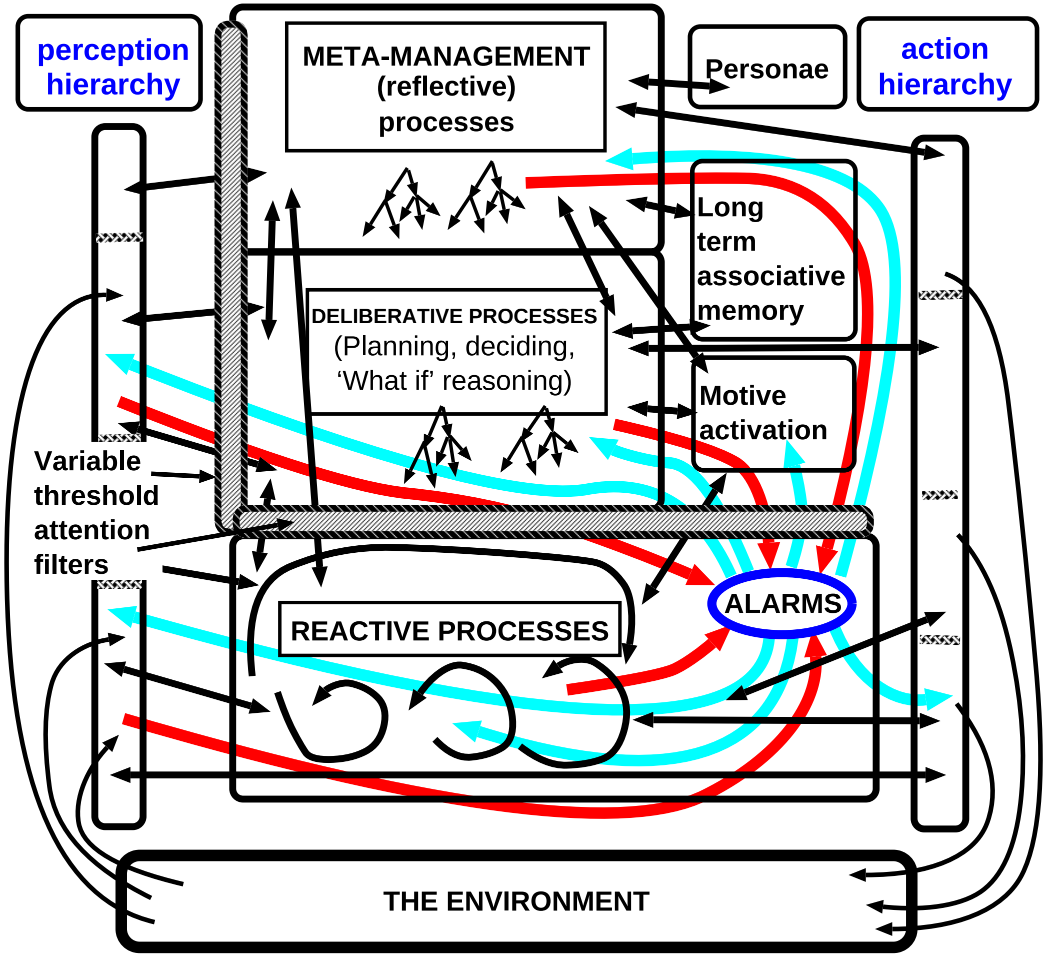
Likewise we should discuss what it is to be emotional (angry, frightened, excited because an old friend is coming home, jealous of a colleague was promoted, etc.) instead of focusing so much on emotions.

This change in style may help us avoid false generalisations, because their reformulations will be patently absurd (e.g. ‘being emotional is a prerequisite for being intelligent’).

A hypothetical Human-like architecture:
H-CogAff (See http://www.cs.bham.ac.uk/research/cogaff/)

This is an instance (or specialised sub-class) of the architectures covered by the generic schema called “CogAff” above.

Many required sub-systems are not shown.


perception hierarchy
META-MANAGEMENT (reflective) processes
Personae
action hierarchy
Long term associative memory
DELIBERATIVE PROCESSES (Planning, deciding, ‘What if’ reasoning)
Motive activation
Variable threshold attention filters
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

Where could it come from?

Various trajectories:

- evolutionary,
- developmental,

  Altricial species build their architectures while interacting with the environment?
- adaptive,
- skills developed through repetition (how?)
- social learning, including changing personae...

(This is an illustration of some recent work on how to combine things: much work remains to be done. This partly overlaps with Minsky’s *Emotion machine* architecture.)

Some notes on H-CogAff and similar things

- The diagram is static, but natural archictures are not static – they develop over time. There is no reason to believe that all the virtual machine components present in a teenager are there in a new-born infant.
- How the architecture grows may differ considerably between precocial species (highly competent at birth or hatching) and altricial species (helpless and cognitively incompetent at birth or hatching).

  The distinction is being explored in papers and talks in the Cogaff project directory, and within the CoSy project

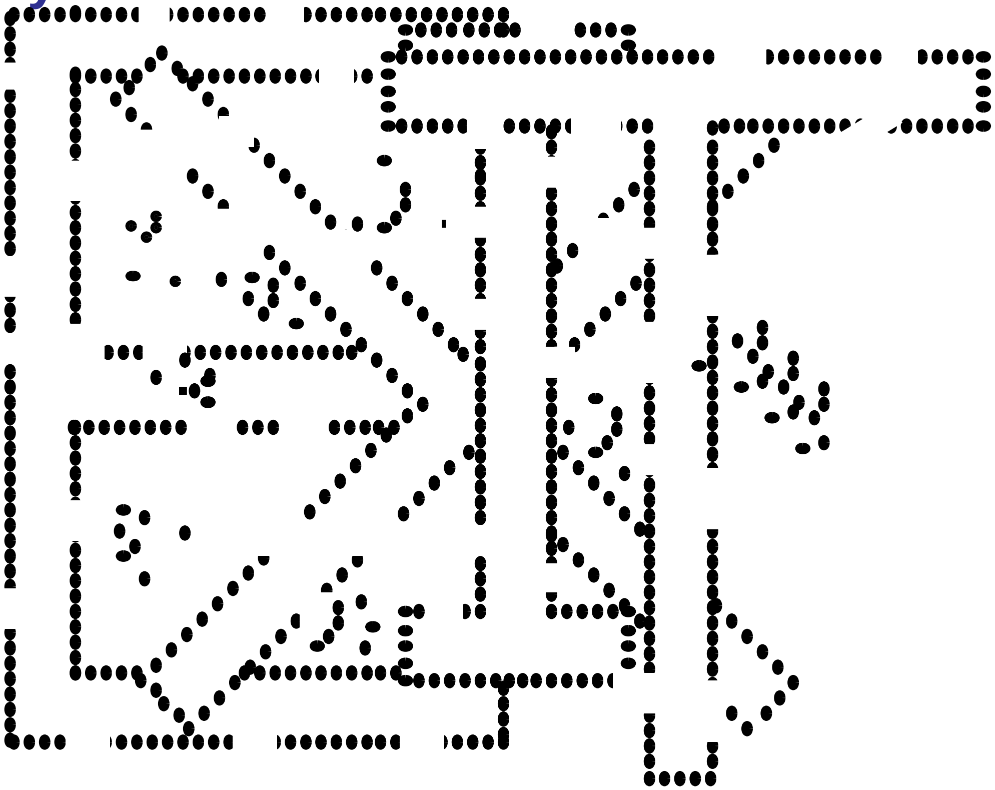
  http://www.cs.bham.ac.uk/research/cogaff/
  http://www.cs.bham.ac.uk/research/cogaff/talks/
  http://www.cs.bham.ac.uk/research/projects/cosy/
  http://www.cs.bham.ac.uk/research/projects/cosy/PlayMate-start.html
- Most computing power is probably in the evolutionarily old reactive mechanisms, though biological implementation of the mechanisms for deliberative and meta-management systems may be very expensive because of their generality and flexibility rather than their power.
- Most people find it hard to understand why perceptual and action mechanisms (in human-like systems and many other animals) need to be layered, with different layers using different ontologies and mechanisms).
  An example from vision follows.

How quickly do you see the word made of dots in the next slide??

Why do human-like systems need concurrent multi-level perception?

(As proposed in H-CogAff)

Answer: In order to cope with rapid recognition of high level structures in complex and messy scenes.



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 (or tasks presented in vision research laboratories)

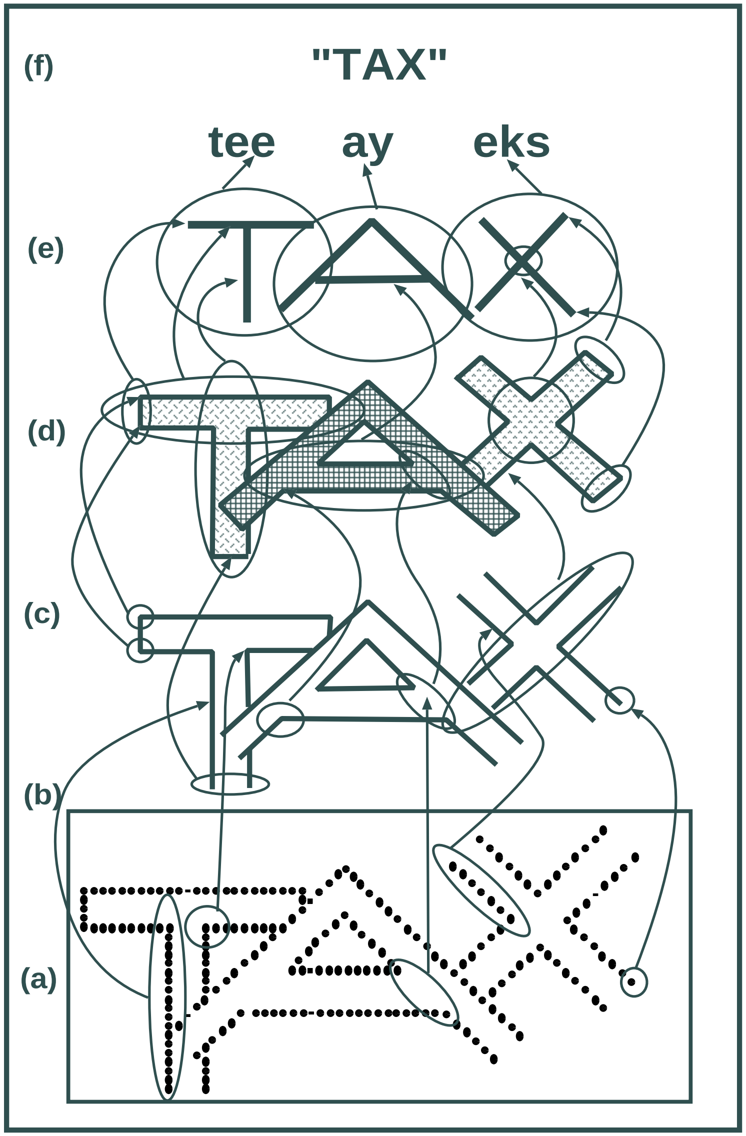
Multiple levels of structure perceived in parallel

Conjecture: Humans process different layers of interpretation in parallel.

Concurrent data-driven and top-down processing helps to constrain search.

Several ontologies are 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 (agglomeration) there is also interpretation, i.e. mapping to a new ontology.
- Concurrent perception at different levels can constrain search dramatically (POPEYE 1978) (This could use a collection of neural nets.)
- Reading text would involve even more layers of abstraction: mapping to morphology, syntax, semantics, world knowledge

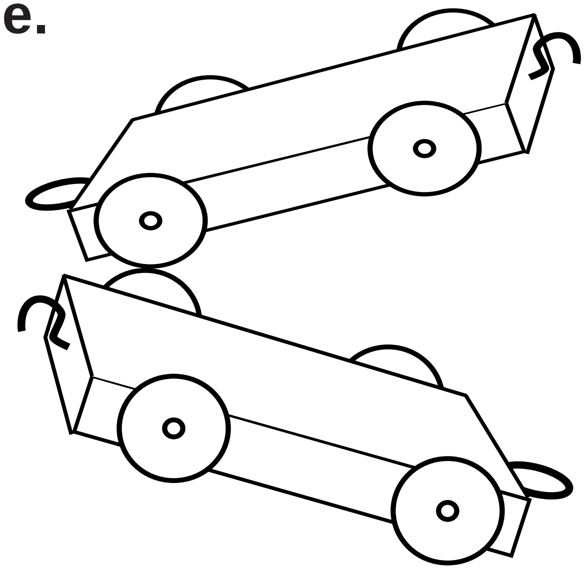


From *The Computer Revolution in Philosophy* (1978)
http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html

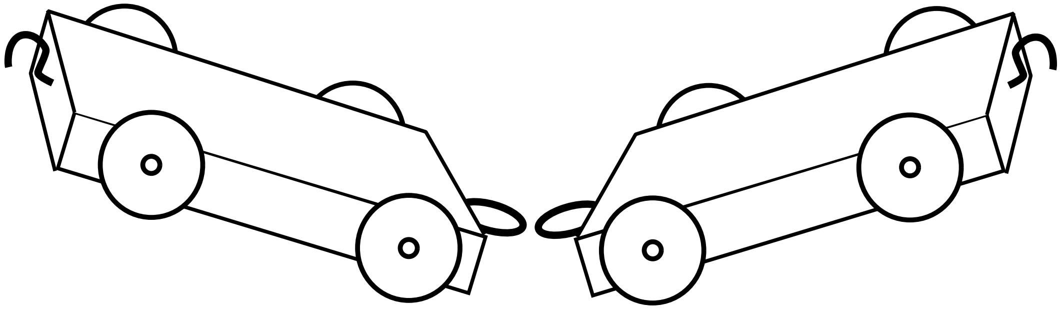
Affordances required for building trains

Perceiving affordances requires high levels of abstraction and representation of what might happen or be done.

How might you move the trucks on the right to join them together?



What capabilities are required in order to see why this will not work?



How might the perceived affordances be represented – in a computer, or in a brain?

What changes between a child not understanding and understanding?

See the video of a child aged about 18 months failing to understand the affordances in two rings as opposed to a ring and a hook:

http://www.cs.bham.ac.uk/˜axs/fig/josh34_0096.mpg

A few weeks later, he seemed to understand. WHAT CHANGED?

Could talking to a child speed up the change? How?

HOW THINGS CAN GO WRONG

Recent discussions of emotions, in AI, psychology and Philosophy, illustrate what can go wrong when science is driven by wishful thinking rather than a disciplined design-based approach.

Are emotions needed for intelligence?

Damasio’s Error

- In 1994 Antonio Damasio, a well known neuroscientist, published his book *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 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 affect had not had its rightful place in psychology, and cognitive science, but the claims were more moderate.

  E.g. a journal called *Cognition and Emotion* was started in 1987.

  At IJCAI in 1981 Monica Croucher and I argued that intelligent robots will have emotions, as side-effects of other things, not that they will need them.

Damasio's examples

Damasio's argument rested heavily 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

- 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/

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

In a nutshell, here is 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 emotional capabilities

P2 Damage to frontal lobes impairs intelligence

C Emotions 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 elsewhere.

Compare this argument

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

P1 Damaging the battery stops the horn working in a car

P2 Damage to the battery prevents the car starting

C A functioning horn is required for the car to start

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. 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.

Very intelligent people keep falling for the argument.

WHY?

E.g. Susan Blackmore did not notice the fallacy when summarising Damasio's theories. See page 285 of her excellent recent book *Consciousness: An Introduction (2003)*. She has now informed me that she agrees that the argument used 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.*

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?

FIRST draft answer:

many cases of being emotional involve having an ‘alarm’ mechanism of the sort mentioned earlier that has been triggered

Example demos

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

They include

- **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/
(Rather large files unfortunately.)

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

They require a linux or unix system, or vmware running on windows to simulate linux.

Varieties of definitions of ‘emotion’

Part of the problem of studying natural phenomena 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’.

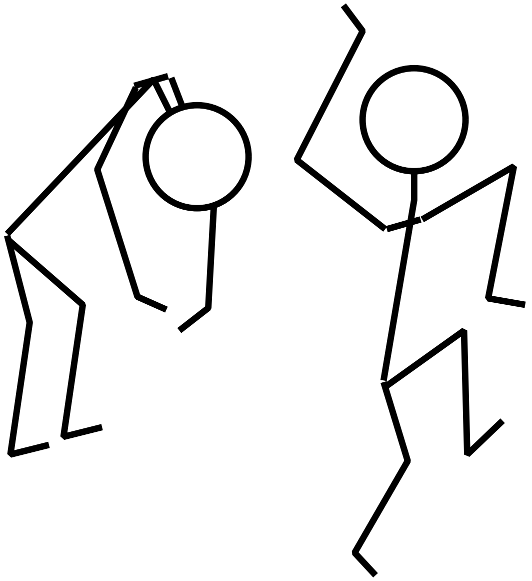
In the psychological literature, for instance, there are attempts to define emotions in terms of

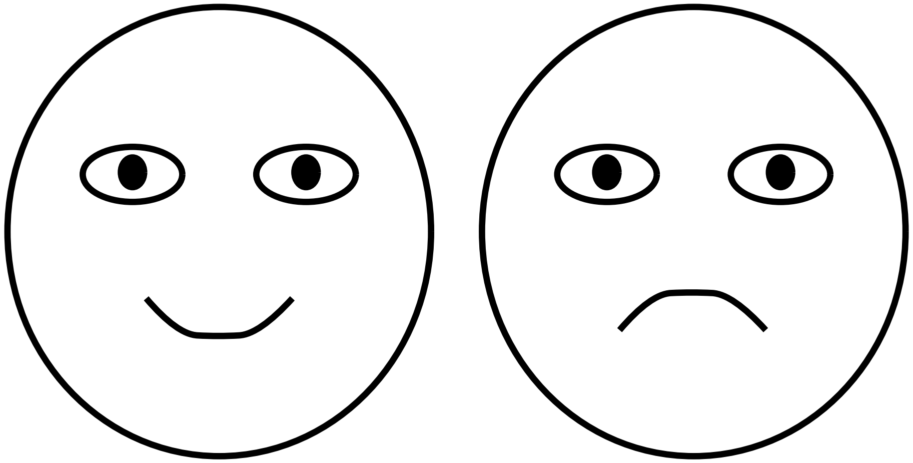
- social relations and interactions between people
- the kinds of states in the environment that produce them
- the kinds of behaviours they produce
- kinds of input-output relations (combining both the above)
- ‘skin-level’ and ‘sub-skin-level’ states and processes, e.g. whether hairs stand on end, 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

.....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.

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 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, each providing its own family of concepts.

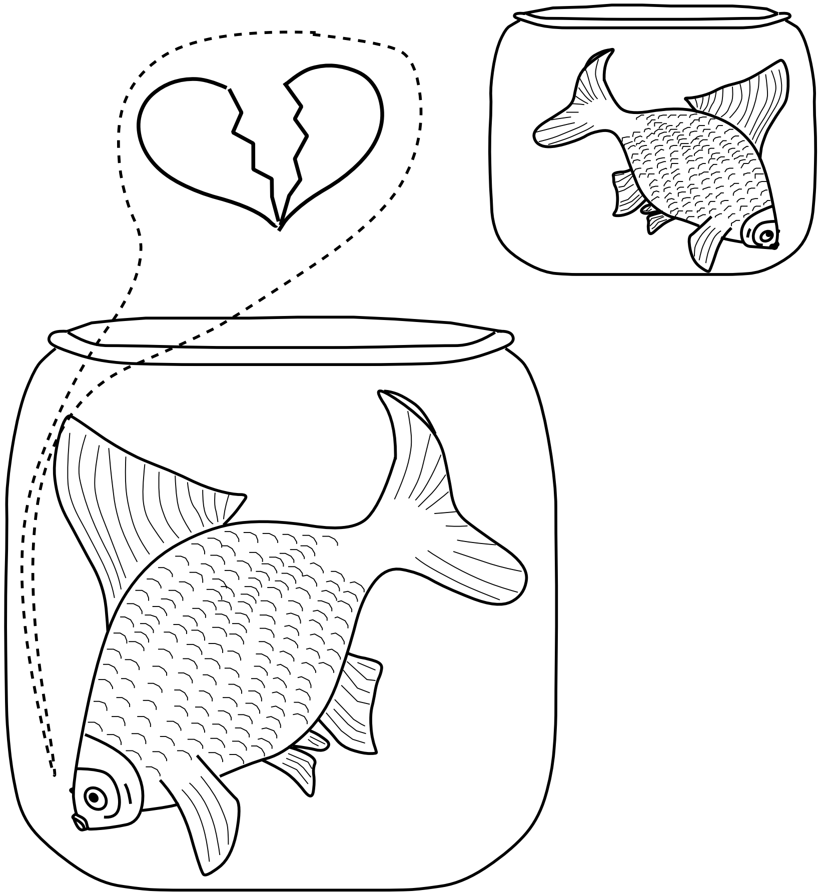
So we need many periodic tables generating different sets of concepts.

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 the 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?

WHY CAN’T A GOLDFISH
LONG FOR ITS MOTHER?



Towards a general framework

We need to talk about “information-using systems” — where “information” has the everyday sense, not the Shannon 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 structures 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.

Examples of types of processes involving information

- Acquisition
- Filtering/selecting
- Transforming/interpreting/disambiguating
- Compressing/generalising/abstracting
- 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.

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 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, 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/#inf

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

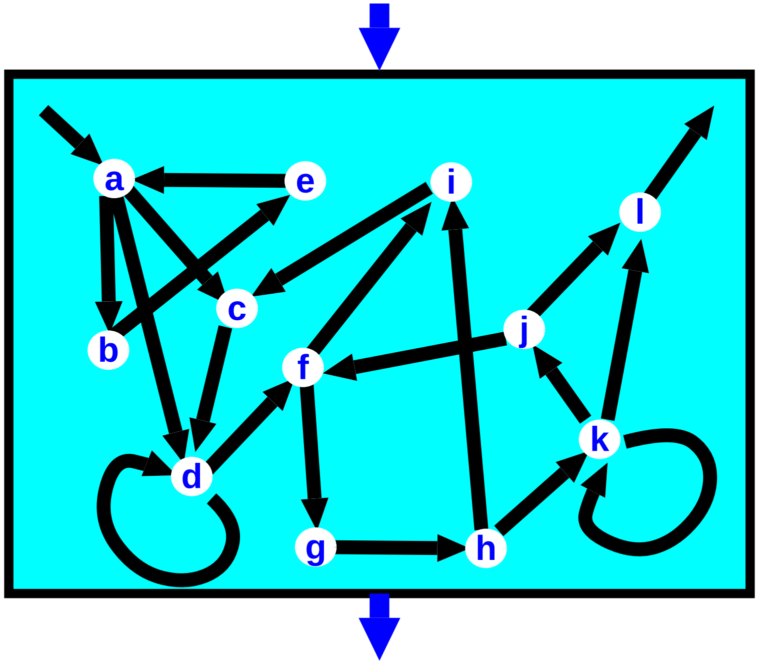
- They can change their structures without having to rebuild the underlying physical machinery
- They can switch between states containing different structures very quickly, far more quickly than physical structures can be reorganised

  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 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.



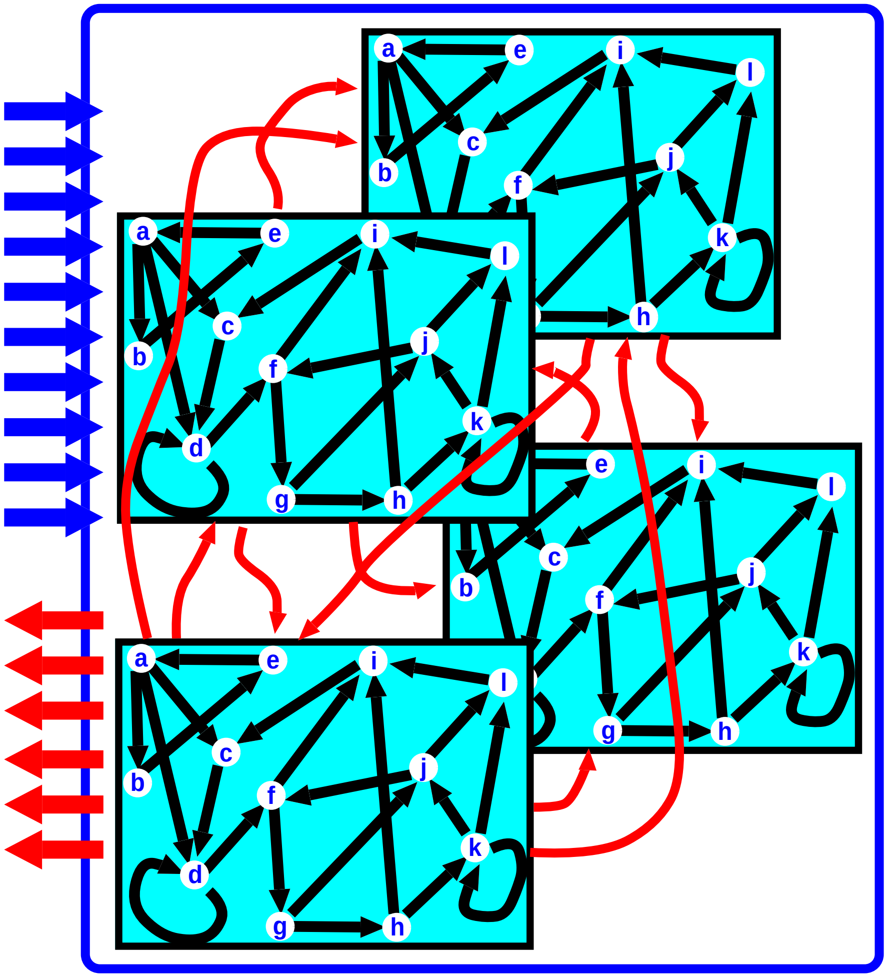
This is how many people think of functionalism: there’s a total state which affects input/output contingencies, and each possible state can be defined by how inputs determine next state and outputs.

(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!).

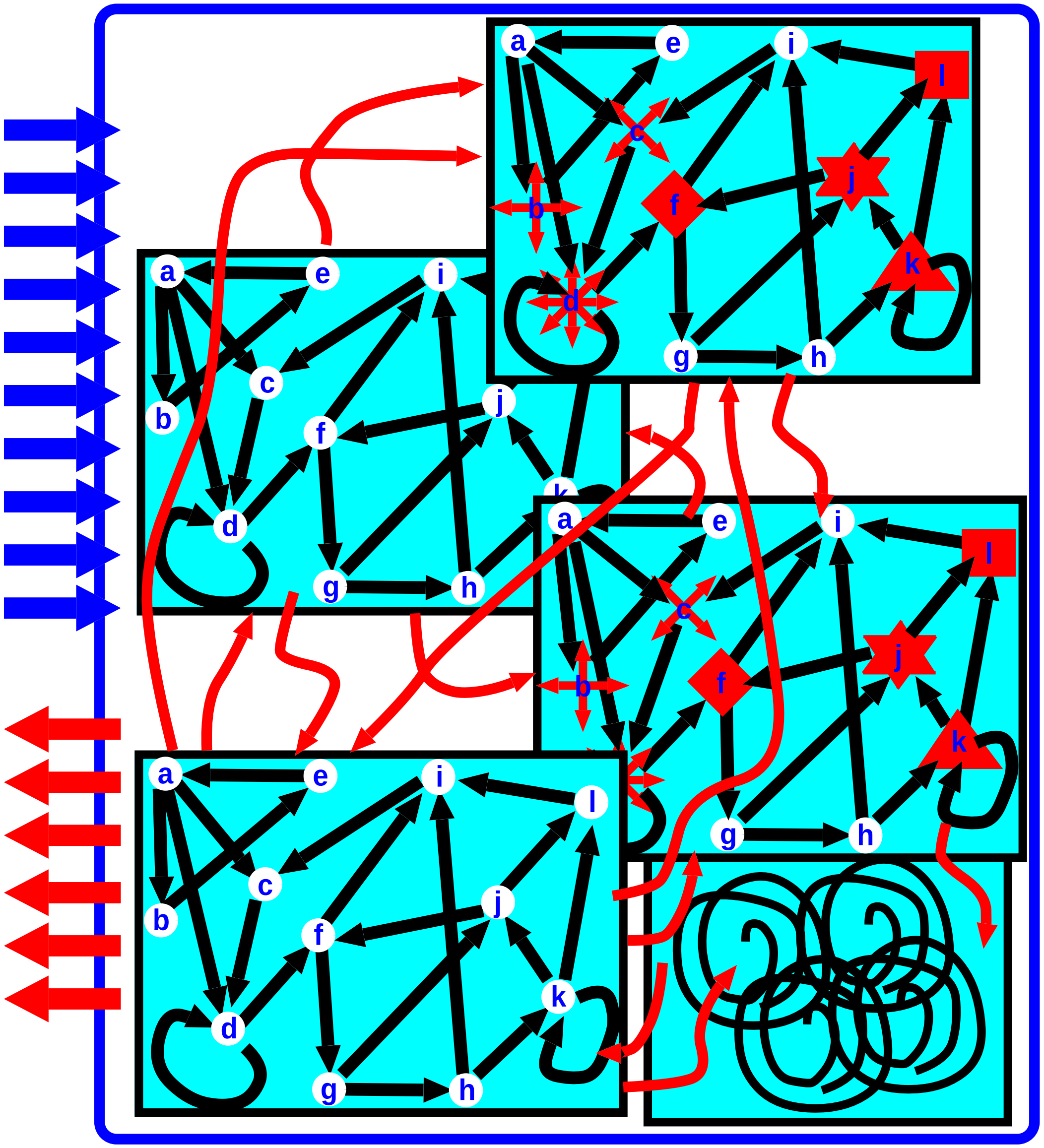
The previous picture is misleading

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/cogaff/talks/#inf
- VMF allows some processes to have the effect of providing control information for others, and for different processes to compete for control.
- If all control is dubbed 'emotional' the label becomes vacuous: but it may be useful to recognize some special cases as emotional, namely some of the cases where where one process disrupts, aborts, suspends, or otherwise interferes with another — e.g. when we are 'moved' by something.

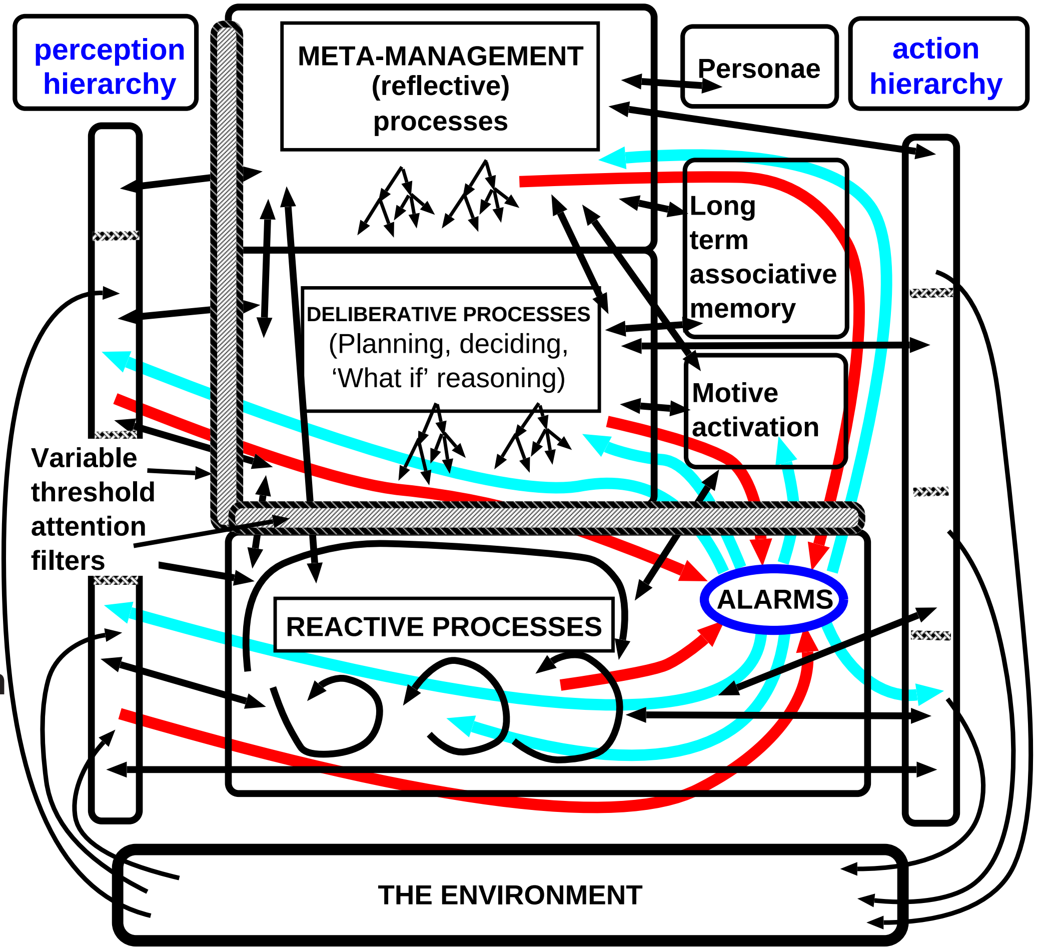
Our own work in Birmingham
The architecture of a human mind

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

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 Cognition & Affect web site 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.


perception hierarchy
META-MANAGEMENT (reflective) processes
Personae
action hierarchy
Long term associative memory
DELIBERATIVE PROCESSES (Planning, deciding, 'What if' reasoning)
Motive activation
Variable threshold attention filters
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

See other Cognition and Affect papers and talks for details

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 much simpler than the 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 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

For lots more on this
SEE 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)

DEMO-MOVIES:

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

SLIDES FOR TALKS:
http://www.cs.bham.ac.uk/˜axs/misc/talks/
http://www.cs.bham.ac.uk/ axs/misc/talks/#talk3

(One of several on emotions)

Analysing requirements — How can a machine:

- See structured but changing (rigid and flexible, inanimate and animate) objects.
- Build plasticine, lego, tinker toy and meccano models – and want to do so.
- Dress and undress dolls, or itself.
- Learn its way round a room, a house, a garden, a village.
- Climb up a chair to fetch a fragile object off a shelf.
- Communicate with other intelligent systems
  - Asking questions and giving answers
  - Requesting or providing explanations, and using them
  - Giving advice or help on how to do something
  - Warning someone who is about to sit on or climb onto a broken chair.
  - Reporting what it did last week.
- Explain how a pendulum clock works.
- Explain what a clock is for.
- Become more skilled at catching a ball, and handing things to people.
- Tie shoelaces.
- Reason about geometrical relations.
- Learn to count, then later use that ability in many tasks,
  as discussed in http://www.cs.bham.ac.uk/research/cogaff/crp/chap8.html
- Learn to think about numbers, then later about infinite sets.
- Feel fear, pity, shame, pride, jealousy, ... enjoy dancing, music, painting, poetry.
- Care about how another feels.
- Help a less able person with everyday tasks?

  E.g. a younger child, or someone physically infirm, or blind, or deaf, or ...

This isn't a project with a well-defined endpoint

COMPARE THE PROJECT OF FINDING A CURE FOR CANCER: THERE ISN'T ONE BUT THERE ARE MANY MILESTONES, INCLUDING CURES FOR PARTICULAR CANCERS.

The task of producing a robot that has all the characteristics of a typical human child, including

- similar physical skills
- similar patterns of learning and development
- similar likes and emotional responses

would require so many advances in so many areas of science and engineering, including materials science and mechanical engineering, that it is unlikely to be achieved in the foreseeable future.

However, a combination of a new subset of child-like abilities based on a deep understanding of how multiple human abilities can be integrated, might be achieved in 15-20 years by a judicious choice of simplifications of human competences.

If such robots can be given useful applications as helpers for blind, or disabled people, as described above, that will be evidence that the simplifications were not a form of 'cheating' to make the goals achievable.

THANKS

The ideas presented here owe a lot to interactions
with Marvin Minsky and Push Singh at MIT,
e.g. see Minsky's draft chapters for *The Emotion Machine*
http://www.media.mit.edu/˜minsky/

and some of the work of John McCarthy,
e.g. his paper on the well-designed child:
http://www-formal.stanford.edu/jmc/child1.html

and many other people...

Thanks also to the developers of Linux
and other free, portable, reliable,
software systems,

e.g. Latex, Tgif, xdvi, ghostscript, Poplog/Pop-11, etc.

For more on this proposal see
http://www.cs.bham.ac.uk/research/cogaff/gc