WHEN WILL REAL ROBOTS BE AS CLEVER AS THE ONES IN THE MOVIES?

Aaron Sloman

A.Sloman@cs.bham.ac.uk

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

School of Computer Science
The University of Birmingham

DAMTP Cambridge

(Department of Applied Mathematics and Theoretical Physics)

21 Feb 2003

Previously presented at ASE Conference
Birmingham, 4th January 2003

These slides are available online here
http://www.cs.bham.ac.uk/research/cogaff/misc/talks/#talk20

They will be revised from time to time.

Software for this presentation

ADVERTISEMENT

This talk was prepared using only
reliable, portable, free software,
e.g. Linux, Latex, ps2pdf, gv, xine,
xdvi, dvips, acroread, Poplog, etc.

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

I am especially grateful to
the developers of Linux for making it
unnecessary to use Windows
on my desktop PC or my portable PC.

Abstract

During the second half of the 20th Century, many Artificial Intelligence researchers made wildly over-optimistic claims about how soon it would be possible to build machines with human-like intelligence. Some even predicted super-human intelligent machines, which might be a wonderful achievement or a disaster, depending on your viewpoint. But we are still nowhere near machines with the general intelligence of a child, or a chimpanzee, or even a squirrel, although many machines easily outperform humans in very narrowly defined tasks, such as playing certain board games, checking mathematical proofs, solving some mathematical problems, solving various design problems, and some factory assembly-line tasks.

This talk attempts to explain why, despite enormous advances in materials science, mechanical and electronic engineering, software engineering and computer power, current robots (and intelligent software systems) are still so limited. The main reason is our failure to understand what the problems are: what collection of capabilities needs to be replicated. We need to understand human and animal minds far better than we do. This requires much deeper understanding of processes such as perception, learning, problem-solving, self-awareness, motivation and self-control. We also need to extend our understanding of possible architectures for information-processing virtual machines. I shall outline some of the less obvious problems, such as problems in characterising the tasks of visual perception, and sketch some ideas for architectures that will be needed to combine a wide variety of human capabilities. This has many implications for the scientific study of humans, and also practical implications, for instance in the teaching of mathematics. It also has profound implications for philosophy of mind.

Note on the title:

My colleague Russell Beale once told me that if students ask him what Artificial Intelligence is he says: “Getting machines to behave like the ones in the movies”

The old, old, goal: build a mechanical man

Read about the legend of the Golem

http://www.atomick.net/fayelevine/pk/golem00.shtml

http://www.ced.appstate.edu/projects/fifthd/legend.html

The tale concerns Rabbi Loew of Prague (born 1513) who built a golem (moulded from clay) to do menial tasks – until the golem grew unhappy about not being able to live like people, became angry and ran away.

The Frankenstein story by Mary Shelley can be read online:

http://www.literature.org/authors/shelley-mary/frankenstein/
http://www.sangfroid.com/frank/

More history

For centuries people have tried to make mechanical toys, looking and behaving like animals or people, in parallel with developments in techniques for making clocks, calculators, musical boxes, mechanical looms and other more or less self-controlled machines.

A MACHINE IS SELF-CONTROLLED OR AUTONOMOUS TO THE EXTENT THAT IT CONTAINS ITS OWN CONTROL INFORMATION EVEN IF THE POWER IS EXTERNAL, E.G. A STEAM-POWERED AUTOMATIC LOOM, OR A HAMBURGER-POWERED CHILD.

MORE RECENT DEVELOPMENTS

The word 'Robot' comes from literature:

Karel Capek's play *R.U.R. (Rossum's Universal Robots)* was written in 1920, premiered in Prague early in 1921,

http://www.uwec.edu/Academic/Curric/jerzdg/RUR/

There have been many films and TV series

- Forbidden planet (Robby the robot)
  http://www.movieprop.com/tvandmovie/reviews/forbiddenplanet.htm
- Dr Who, Cybermen, and the robot dog: K9.
  http://www.dwguide.demon.co.uk/
- Star Wars (R2D2, C3PIO)
- Star Trek (hologram robots)
- Blade Runner
- The Forbin Project (Colossus - super intelligent machine)
  http://www.allscifi.com/Topics/Info_8636.asp

and many more

Film makers and story tellers have the luxury of 'make-believe'. They don't really have to analyse the problems of designing a robot that works — they don't have to build one.

There are many techniques for creating the *illusion* of a working human-like robot (including putting a person inside a machine).

What about AI?

- Artificial Intelligence (AI) researchers have tried for decades to produce intelligent machines (like those in movies).
- So far, despite many impressive achievements, they are nowhere near producing a robot with the mind of a five year old child.
- Current robots have general intelligence far below that of a young child, or even a squirrel, or crow, even though some of them may be very good at performing some narrowly circumscribed set of tasks.

For examples, see the Birmingham University Robot page

http://www.cs.bham.ac.uk/research/robotics/cbbc/

the Honda Asimo robot page

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

(Including some movies showing the robots doing things.)

and the Sony AIBO page

http://www.aibo.com/

There's a lot more information at the AITOPICS web site

http://www.aaai.org/AITopics/html/robots.html

Impressive robots made by Honda and Sony

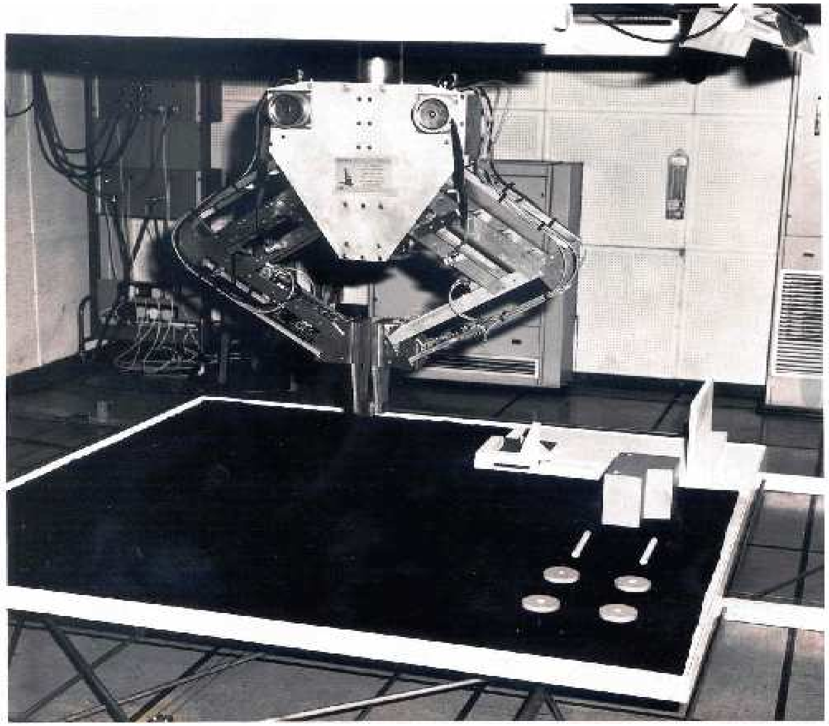
THE STATE OF THE ART IN 2002

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 chimp or squirrel. Mostly they have purely reactive behaviours, lacking the deliberative ability to think or wonder 'what would happen if...'. They also have very little self-knowledge or self-understanding, e.g. about their limitations.

Compare Freddy the 1973 Edinburgh Robot



Some people might say that apart from the 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 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.

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, not psychologists.

Why are current robots still so limited?

1. **Because mechanical engineering and materials science are still far behind what biological evolution has produced (e.g. power-weight ratios, sizes of functional components).**
2. **Because the computing power required to match animal brains, along with constraints of size, weight, energy consumption, etc. are so great.**
3. **ABOVE ALL BECAUSE WE DON'T YET KNOW WHAT TO DO TO REPLICATE ANIMAL INTELLIGENCE.**

   I.e. We don't yet have a deep understanding of what animal intelligence includes.

   In engineering terms: we don't know the functional requirements for the designs we are trying to produce – we don't really know what sort of software is required, what it is supposed to do, let alone how to design and implement software to meet the requirements.
4. ***This point is missed by people who proclaim that because computers are becoming much more powerful very rapidly we shall soon be able to produce super-intelligent machines.***

   That view assumes that we shall know how to use all the new power.

N.B. I AM NOT SAYING THAT THE TASK IS IMPOSSIBLE

Compare, J.R. Lucas, John Searle, Roger Penrose, ...

How should we describe the task?

- Merely saying that we want to build machines with human-like (or animal-like) capabilities assumes that we know what those capabilities are – whereas we don't as yet

  although we are learning, partly through doing AI, finding how un-human-like our systems turn out to be, and then studying the reasons for the failures.

- Making progress requires a meta-level theory of what we need to know in order to specify those capabilities, to help us do the research to discover the capabilities, so that we can then try to design systems that have them.
  - In part this requires us to find the right way to describe *the environment*: which is different for different organisms, even in the same physical location.
  - The correct description of the environment depends on the organism, i.e. on the design.
  - There is a circular bootstrapping process, in which doing AI helps us understand what the task is, by analysing the unexpected inadequacies of our requirement specifications and designs e.g. in 'field tests'.

    (This process does not conform to standard models of software engineering, though it is part of the reality of software engineering, especially the process of designing systems to interact with humans.)

- We also need a way to survey the space of possible designs for intelligent entities, so that we can understand alternative options available and see how humans are related to other organisms and machines.

What is the space of possible designs?

Despite enormous progress in the design of computers, programming languages, operating systems and many applications of computing technology our understanding of the space of possible information processing systems is still in its infancy.

We still have only limited understanding of

- **Possible ways of encoding information (including continuously varying and structurally changing information)**
- **Possible ways of manipulating information**
- **Possible ways of applying information for practical purposes**
- **Possible ways of combining components into larger integrated systems**
- **Possible ways of controlling complex multi-functional systems**
- **Possible physical mechanisms for implementing the above**

We need a meta-theory to help us with the task of exploring types of solutions.

The CogAff framework sketched below is a first draft attempt to provide part of such a meta-theory (only a part).

Within the CogAff framework we can explore a variety of very different information-processing architectures including those required for insects and a tentatively proposed H-CogAff model for human-like systems.

Note on 'Information'

- Often I am asked what I mean by "information" and "information-processing"?
- The full answer is quite complex.
- Partial answers can be found in talk 4 and talk 6 here:
  http://www.cs.bham.ac.uk/research/cogaff/talks/

To a first approximation, when you know

- the forms that information can take,
- the variety of contents it can have,
- the various ways it can be
  - acquired,
  - manipulated,
  - analysed,
  - interpreted,
  - stored,
  - transmitted,
  - tested,

  and, above all,
  - used,

then, to that extent, you know what information is.

That knowledge grows over time, like our knowledge of what energy is.
Did Newton understand "energy" as we do?

Understanding an information-processing system

A designer of a working information-processing system, or someone trying to understand such a system, requires knowledge about the following:

1. what the parts of the system are, and possibly how they are designed

   (These are not necessarily *physical* parts – e.g. the scheduler, memory manager, file-system manager, in a computer operating system – are parts of a *virtual machine*.)

   Understanding of a system may go down to a certain level, which is taken for granted.

   Some of the parts will contain symbols or other structures that express various kinds of information for the system. For instance, some parts may have information about other parts, as in an operating system. Some will have information about the environment.

2. the relationships between the parts, including structural, causal, semantic, and functional relationships

   Functional relations are (roughly) causal relationships that contribute to some need, goal, or purpose, e.g. preserving the system, or optimising its response times.

3. the portion of the environment with which the system interacts, and the structural, causal, semantic, and functional relations between parts of the system and its environment.

   (Different parts may interact concurrently with the environment.)

These are all aspects of the *architecture* of the system: some are intrinsic aspects, while others are extrinsic (environmental).

These aspects need to be understood both by designers of systems and by scientists studying such systems (e.g. psychologists.)

Show demos

Here are examples of an amazing animal, and some not so amazing computer programs indicating some of the requirements:

- Betty the crow (online video and audio interviews available):
  http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm
  (We'll analyse some of the requirements for Betty's task later.)
- Simulated Sheepdog demo
  An example of a purely reactive system that gives the impression of being able to think ahead, i.e. the simulated sheepdog displays apparent deliberative behaviour as it fetches sheep to the front of the pen and then drives them in, going round obstacles where necessary.

  But it is purely reactive, though some of the reactions change *internal* states.

  The code for the sheepdog demo (using our SimAgent toolkit) is here
  http://www.cs.bham.ac.uk/research/poplog/newkit/sim/teach/sim_sheepdog.p
  It needs Poplog http://www.cs.bham.ac.uk/research/poplog/freepoplog.html
- SHRDLU (Based on work by T.Winograd at MIT in 1971!)
  This is a deliberative program: it can construct complex possible interpretations of a sentence or possible plans, compare them and select the best.

  The code for it is part of the Poplog RCLIB package
  http://www.cs.bham.ac.uk/research/poplog/freepoplog.html#rclibfiles
- Play with an online purely reactive (partly random) Eliza chat-bot, inspired by Weizenbaum's 1966 Eliza: http://www.cs.bham.ac.uk/research/poplog/eliza/eliza.php

Digression: Educational Conjecture

Conjectures:

- *Teaching school-children to design, implement, analyse, and document programs like those I have demonstrated will be of more benefit in the development of young minds than teaching them to use word processors or web browsers, as happens now – since teaching students to use computer programs without ever learning to design and develop programs turns 'couch potatoes' with 'mouse potatoes'. All that mouse-clicking looks active but is often essentially passive.*
- *Teaching AI programming, especially if natural language processing is included (like SHRDLU and Eliza) could also get more girls interested in learning to program – helping to stretch their minds in deep new ways – e.g. helping them to think about human minds from the standpoint of software engineers, rather than just the standpoint of novelists, poets and movie-directors.*
- *Education should be based on The Five Rs: reading, writing, arithmetic and programming (including specifying, designing, testing, debugging, explaining, and documenting working systems)*

Returning to our main theme: in order to make progress with robots that are closer to human and animal intelligence, we need to analyse the aims of AI – and the difficulties.

CAN MACHINES HAVE MINDS? including percepts, feelings, emotions, ...?

ONE (PARTIAL) ANSWER:

It is obvious that machines can have minds and be intelligent, because

- Humans are machines and they have minds and intelligence
- Likewise many other animals, e.g. chimpanzees, elephants, dolphins, dogs, crows, ... have mental states and processes.

So that leads us to rephrase our questions:

- What sorts of machines are there,
- What sorts of minds can different sorts of machines have?
- What sorts of intelligence can the different sorts of machines have?
- What sorts of percepts, feelings, emotions, consciousness, can the different sorts of machines have, e.g. humans, elephants, dolphins, mice, fleas, robots of various kinds?

- One of the roles of AI is to investigate these questions.
- Another is to help us build new kinds of useful machines.

AI has two main kinds of goals

- **Science** (understanding)

  i.e. studying things that already exist or might exist, explaining how they work, searching for general principles relevant to understanding
  - people,
  - other animals,
  - intelligent machines of the future,
  - and perhaps creatures from other parts of the universe.

- **Engineering** (making, fixing, preserving, changing)

  i.e. using that knowledge to solve practical problems, including
  - making new useful kinds of machines,
  - producing new forms of entertainment
  - perhaps helping us manage ourselves better,
    e.g. in education, therapy, ...
    - because we understand ourselves better
    - because we have new tools.

There's more on what AI is, with pointers to sources of information, here
http://www.cs.bham.ac.uk/~axs/misc/aiforschools.html

Physical and virtual components, relations etc.

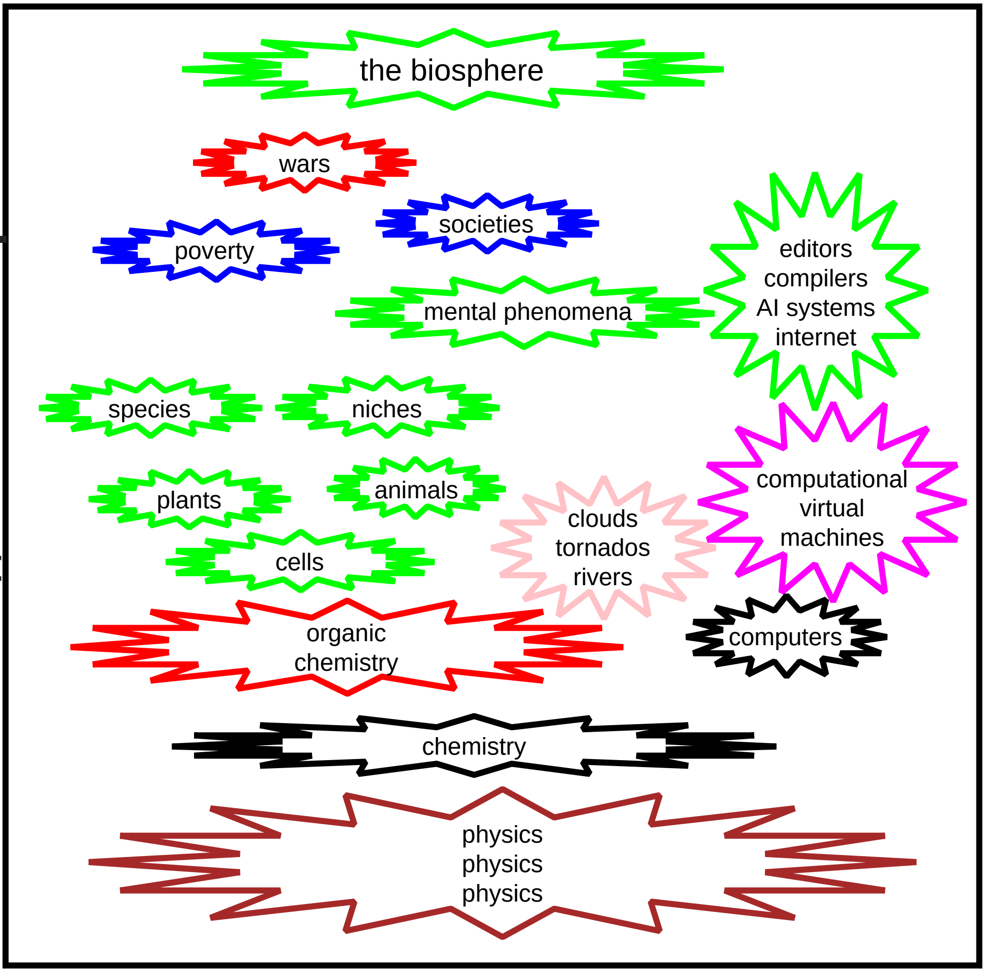
When we talk about components, inputs, outputs, causal interactions, etc. we are referring to phenomena that exist at various levels of abstraction, including components of virtual machines.

- **The components that we are interested in are not just physical components.**
  (They may include parsers, compilers, tables, graphs, schedulers, image interpreters...)
- **The various kinds of relations, properties, dynamical laws are not restricted to those investigated in the physical sciences**
  (i.e. we have to go beyond physics, chemistry, astronomy, geology,... e.g. we have to include relations like *referring to*, *implying*, *monitoring*, *evaluating*, *selecting*, etc.)
- **We have to understand virtual machines at various levels of abstraction. This includes understanding how virtual machines interact with the physical world.**

  For example, when a chess-playing program runs on a computer, the chess virtual machine includes entities and relationships like: kings, queens, pawns, rows, columns, colours, threats, moves of a piece, etc.
  These are not things that a physicist or chemist or electronic engineer can observe by opening up the machine and measuring things.
  Software engineers design, implement and debug virtual machines.
  Many people use virtual machines without realising that they do.

  N.B.: Action-selection in a *virtual* machine can cause changes in *physical* parts or in external physical objects.

Ontological levels are everywhere



At all levels there are objects, properties, relations, structures, mechanisms, states, events, processes and CAUSAL INTERACTIONS.

E.g. poverty can cause crime.

But they are all ultimately realised (implemented) in physical systems.

Different disciplines use different approaches to study these levels (not always good ones).

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

Our emphasis on virtual machines in organisms is just a special case of the general need to describe and explain virtual machines in our world, e.g. economic machines..

See our IJCAI'01 Philosophy of AI tutorial for more on levels and causation:
http://www.cs.bham.ac.uk/˜axs/ijcai01/

Information-processing machines

Both the scientific and the engineering goals of AI
are concerned with information-processing machines:
how to understand them and how to build them.

This makes AI very different from
the study of other kinds of machines,
e.g. energy-transforming machines
and matter-manipulating machines
as studied in physics, chemistry
and older kinds of engineering

Information processing machines can be virtual machines
as abstract as the information they are processing...

Information-processing machines, including virtual machines, have existed for millions of years, but only recently have we begun to understand how to study them and how to design and build them – apart from the very simplest types.

THERE IS STILL MUCH THAT WE DON'T UNDERSTAND.

Studying the environment (the niche) is very important, and often very hard

It may be hard to find out what the relevant environment of an organism is without understanding the organism.

- If an object O has rich interactions with environment ENV (e.g. perceiving, acting, learning, communication) then understanding O requires understanding ENV .
- A special case of ENV : in order to understand a biological organism, one has to understand its *niche*.
- But what that niche is may be far from obvious, as we'll see.
- In particular, besides physical properties, a niche, or environment, for an information processing system O may include what Gibson called "affordances" for O . (J.J. Gibson, *The Ecological Approach to Visual Perception*, Erlbaum, 1986)
- The same physical environment may have different affordances for different organisms, or robots: one physical environment may support many different ENV s for different O s.
- E.g., representing the environment as a vector of measurements may fail to address the features of ENV that O perceives and uses – e.g. the perceived structure of a tree, or an animal, or a partly built nest.

So if O is an information-processing system, then understanding ENV requires understanding O and *vice versa*.

Bootstrapping is required!

You cannot understand what the environment is until you have understood what the organism can perceive, learn, reason about, consider for actions.

You cannot understand what the organism can perceive, learn, reason about, consider for actions, until you have understood what the environment (niche) for that organism is.

For an organism that develops, the environment develops also.

Living organisms as information processors

What is special about living organisms?

- For most physical objects (e.g. a marble rolling down a helter-skelter) there is no clear separation between impinging forces and information used.

  The energy producing motion of such an object comes from the forces determining what the motion (gravity, friction, collision forces, etc.) should be.

  Moreover, the effects are generally direct and instantaneous.

  (Exceptions are cases involving long-distance transmission and thresholds to be overcome, e.g. volcanoes that erupt, or dams that collapse, only after build-up of pressure, etc.)

- Living things separate
  - acquisition of energy
  - acquisition of information

  Both can be stored for future used.

  - Consumption of food provides energy for use later.
  - Much information is stored for use later.

  In the simplest living things information from the environment is not stored for future use, but used immediately.
  Nevertheless it does not always directly drive the “motors”. Rather information is used to switch on and direct internal energy supplies.

  Compare von Neumann on ‘switching organs’ in his 1951 paper.

Information and energy in living organisms (intelligent systems)

In such systems:

- Sensors obtain information, on the basis of which (together with previously stored information) actions are selected. (Both external and internal actions).
- Previously stored energy (mainly chemical energy) provides the forces required to perform the actions.

Some evolutionary developments make the information-processing more and more complex, flexible, varied, and powerful. This increases the energy required to support internal information processing mechanisms – e.g. a large content-addressable memory. This is characteristic of deliberative capabilities implemented in biological mechanisms: they are ‘costly’ to produce and to run.

The result is a type of organism that has to be high up a food pyramid, and is expensive to produce (so parents produce few offspring) and therefore costly (for the gene pool) to lose (so that parents invest a lot of care in offspring).

There cannot be many such species: the vast majority of species have numerous, low cost, relatively unintelligent, individuals.

Compare Pianka’s notion of the r - K selection trade-off.
http://uts.cc.utexas.edu/ varanus/rKcit.html

An evolution-based meta-theory

Thinking about evolutionary stages can help us develop a meta-theory suggesting ways of thinking about possible sorts of designs, and associated 'niches', which

- **provide pressures to produce or modify the designs**
- **provide criteria for evaluating the designs**
- **are themselves modified by designs in co-evolving species.**

  This leads to the notion of a very complex dynamical system of co-evolving, mutually influencing, collections of species and niches, with multiple feedback loops linking different sorts of trajectories.

The processes of change and development involve

- **trajectories in design space,**
- **trajectories in niche space,**

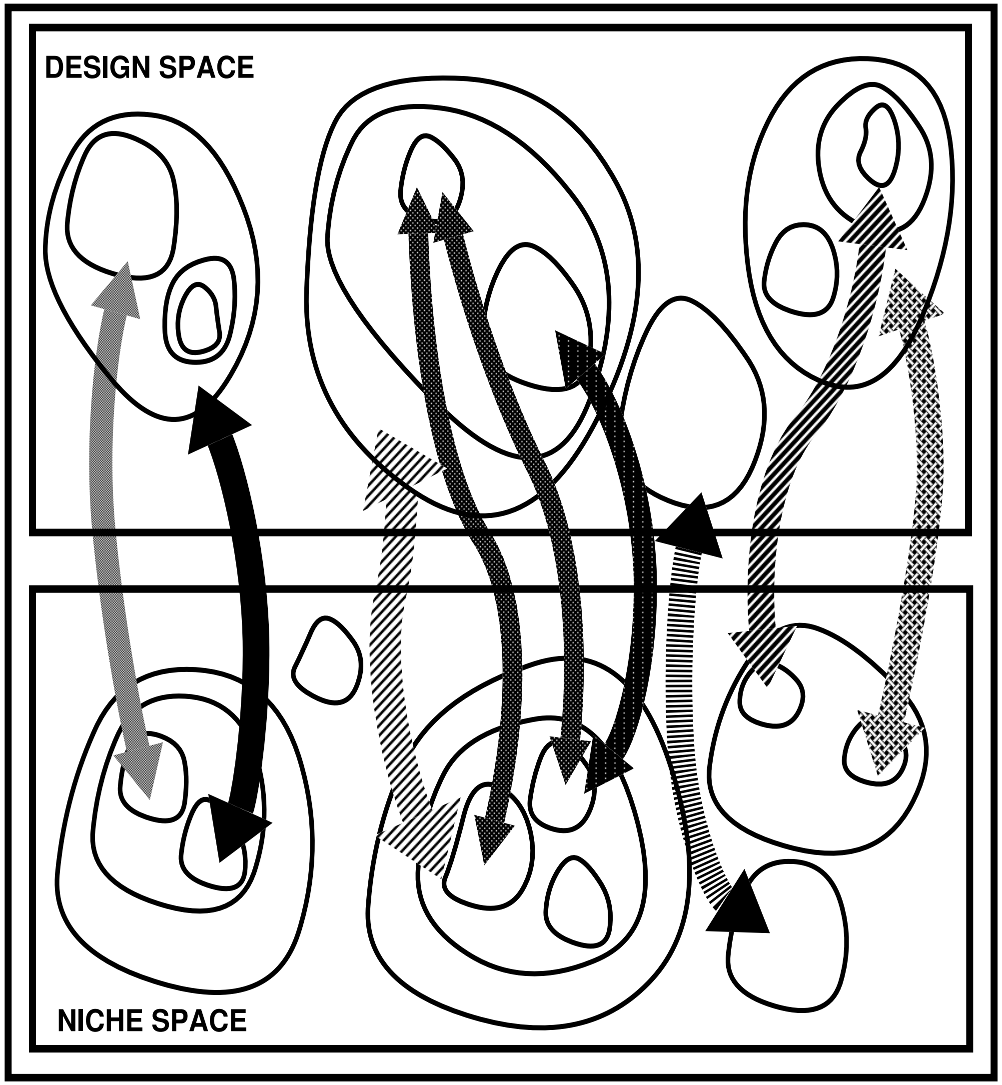
  and

- **feedback between them,**
- **where the trajectories can include both continuous and discrete changes, including structural changes.**

Can we hope to formalise and analyse such dynamical systems mathematically, or will be at best be able to simulate them? How crudely?

Design space and niche space

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



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

Instead we should investigate many (large and small) discontinuities as features are added or removed.

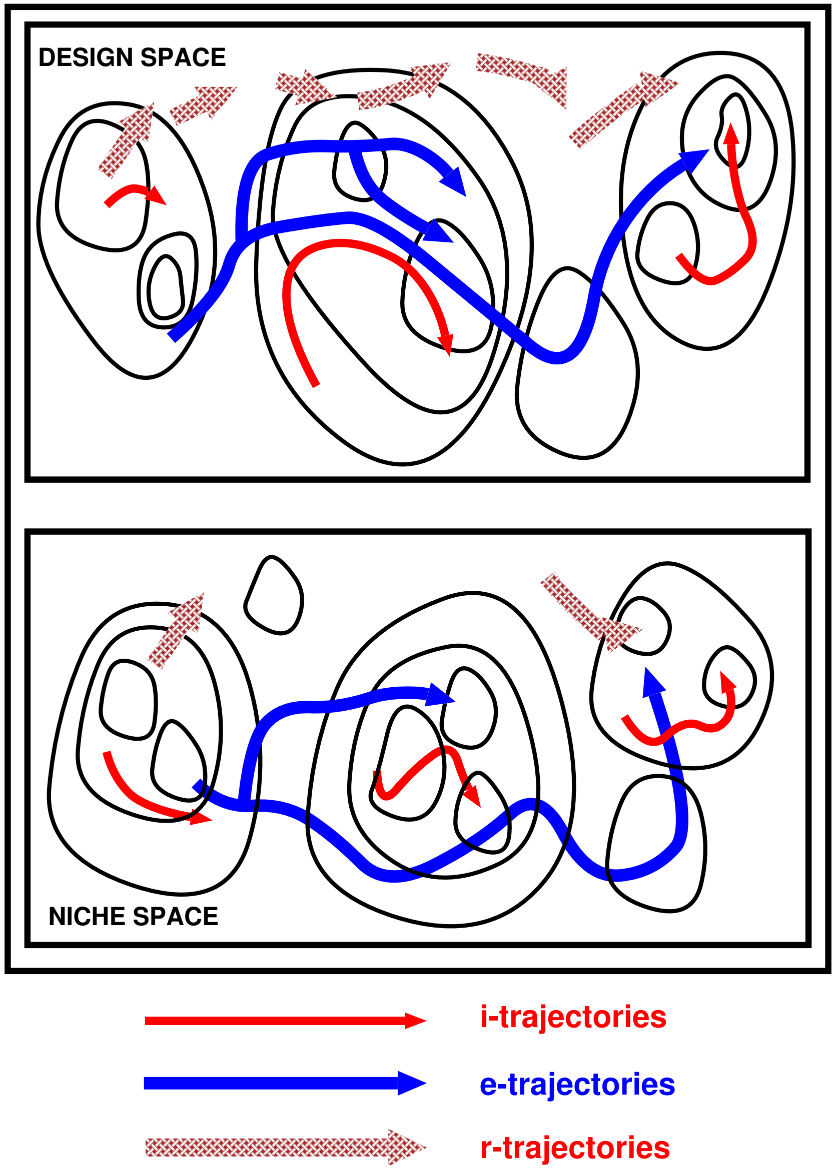
Different sorts of arrows represent different fitness relationships between designs and niches: don't think of fitness functions producing a single value for each design (or an ordering of designs)
The relations differ in kind: they are not simply numerical values, but can include structural descriptions, vectors of values, etc.

Another way of putting this: instead of a niche (set of requirements) pushing whole designs up a fitness slope, it can simultaneously apply distinct pressures on separate components of the design: sensors, motors, memories, internal mechanisms, forms of representation, etc. etc.

Trajectories in design space and niche space

There are different sorts of trajectories in both spaces:

- **i-trajectories:**
  Individual learning and development
- **e-trajectories:**
  Evolutionary development, across generations, of a species.
- **r-trajectories:**
  Repair trajectories: an external agent replaces, repairs or adds some new feature. The process may temporarily disable the thing being repaired or modified. It may then jump to a new part of design space and niche space.
- **s-trajectories:**
  Trajectories of social systems.
- **c-trajectories:**
  e-trajectories that are influenced by cognitive processes (e.g. mate-selection, explicit use of birth control, etc.).



Need for new ways of studying dynamics

- The type of evolutionary process described here includes feedback loops involving multiple discontinuous trajectories in at least two different spaces.
- Do we have the right conceptual tools to study the dynamics?
  - E.g. we need to clarify the nature of different kinds of states: can affective states and non-affective states be distinguished in a principled way.
  - What are cognitive states?
- Can we understand the full variety of ways in which information can be encoded (in physical and in virtual machines), manipulated and used.
- Will we need new mathematics? (The limits of equations.)
- If mathematical methods of analysis are not available, will predictive or explanatory simulation be possible?
  (Perhaps not, if fine details of the total physical environment are all relevant.)

In part the answer may be that we can can only analyse or simulate processes restricted to small regions of design space and niche space – e.g. see the experiments of Matthias Scheutz (http://www.nd.edu/~mscheutz/).

A COROLLARY:
Supposed dichotomies become complex taxonomies.

EXAMPLES:

- We may think we understand what a ‘reactive’ system is, but when we investigate closely we find a space of types of reactive systems with importantly different properties – especially if they are not state free.
- Likewise the category of ‘deliberative’ systems divides into sub-categories when we study different architectures. (Many other examples.)
- There’s also the supposed distinction between things with and things without consciousness. The design-standpoint leads to a discrete space of designs with many intermediate cases between microbes and humans.

A continuum is not the only alternative to a dichotomy

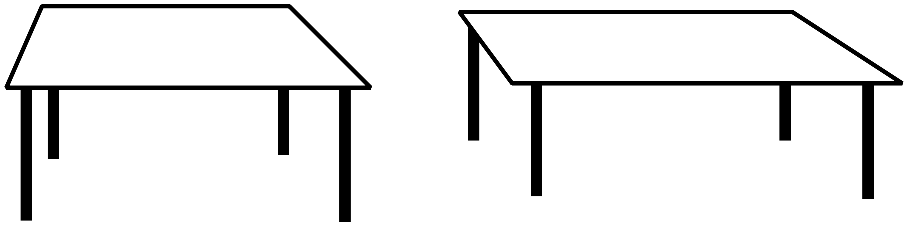
Talking about “degrees of consciousness” often expresses a failure to understand this.

Examples of research problems in AI

Vision – perhaps the hardest problem in AI

How do we get from 2-D patterns of illumination on our retinas to percepts of a 3-D world:

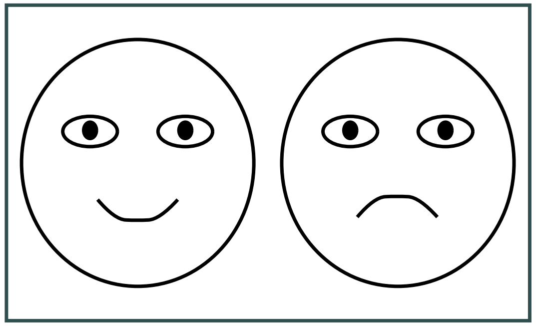
That was done long ago, for simple cases, like tables.



How do we see these as two views of the same 3-D object

How do we see expressions of emotion in faces?

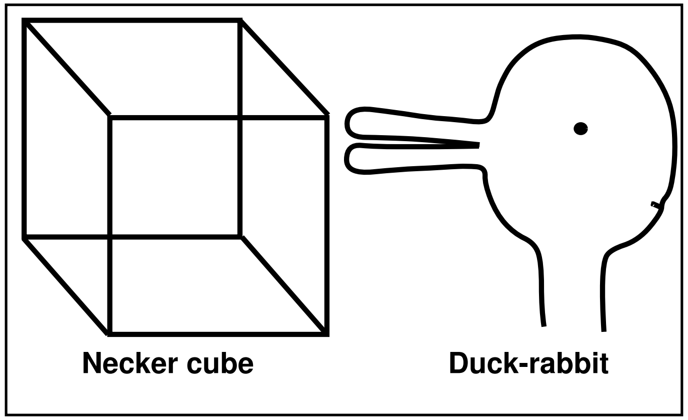
Or in postures, gestures, etc.?



What sort of machine can see eyes as happy or sad?

How can we see the same 2-D visual input in different ways?

When the cube flips, the perceived change is purely geometrical. When the duck-rabbit flips, far more subtle things change. (What things?)



There are many more things to explain, including: perceiving motion, seeing how something works, experiencing visual pleasure, etc.

Most of what we see is much more abstract than physical objects and their physical properties and relationships.

There is far more to perception than detecting what exists in the environment

Betty the crow had to perceive not only the things that were before her at the start:

- The large transparent tube
- The bucket of food in the tube
- The piece of wire

She also had to see the possibility of things that did not exist but might exist, e.g.

- The possibility of the bucket moving up the tube,
- The possibility of the wire being bent and holding its shape
- The possibility of various steps in the process of bending the wire
- The possibility of using the bent wire (which does not yet exist) to lift the bucket of food.

These are all cases of the perception of affordances, whose importance was noted by the psychologist J.J.Gibson.

Affordances are the *possibilities for* and *constraints on* action and change in a situation.

Affordances in an environment depend on the goals and action capabilities of the organism (or robot) perceiving the environment.

An exercise

Imagine you are a crow or a magpie, and you need to build a nest among branches high in trees, using only naturally available materials, such as twigs, leaves, and pieces of grass.

The nest must be rigid enough to survive moderately strong winds and capable of keeping a few eggs safely in place, and later on a few hatchlings.

The only way you can assemble the nest involves using your beak, and occasionally claws, to hold and manipulate the materials, found and fetched by flying around in the neighbourhood.

What are the kinds of affordances you would need to perceive at different stages in the process of finding and selecting the materials, bringing them back, starting off the nest and later adding to it, then deciding when to stop?

I DON'T THINK ANY ROBOTS WILL BE DOING THIS SOON.

Conjecture regarding perceptual layers

It seems that the perception of abstractions, including affordances, is related to the development of central processes that can both make use of the more abstract kinds of information and also help to control the perceptual processing, e.g. by providing context for interpreting ambiguous images.

In other words:

- **Different "layers of" perceptual mechanisms,**
  - **operating concurrently on the same sensory input,**
  - **but producing different kinds of perceptual information,**
- **co-evolve with different *central processing layers* performing different tasks,**
- **and perhaps also with different *layers in a hierarchical action system*.**

This is illustrated below, and provides background for the proposed meta-theory (CogAff schema) for describing information processing architectures.

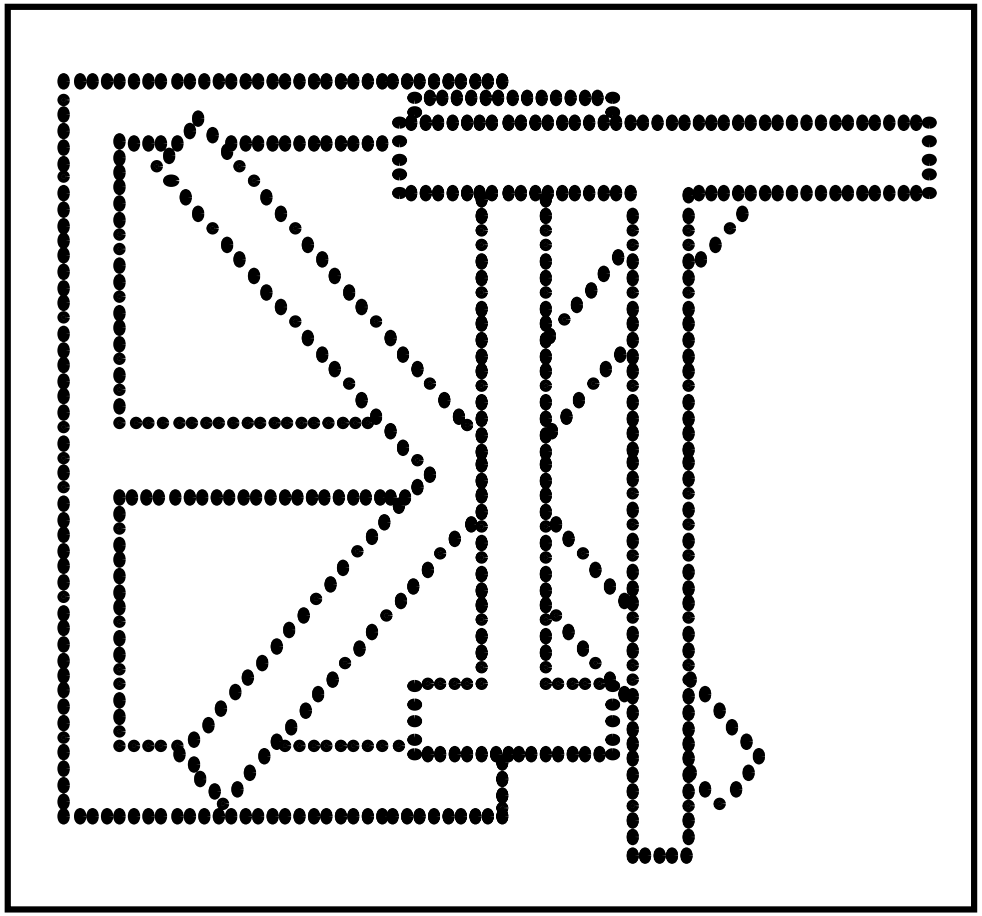
Perceiving meaning in clutter: layered meanings

Humans (and perhaps other animals) are able to see structure and meaning in very cluttered scenes, where structure exists at different levels of abstraction.

People can cope with pictures like this with varying degrees of clutter and with varying amounts of positive and negative noise.

Human performance degrades gracefully, and we often recognize the word before the individual letters have been recognized.

HOW?



See: *The Computer Revolution In Philosophy* (1978) Chapter 9

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

Question:

Does a new-born human infant have a visual architecture able to do this?

Conjecture:
We perceive fragments at different abstraction levels

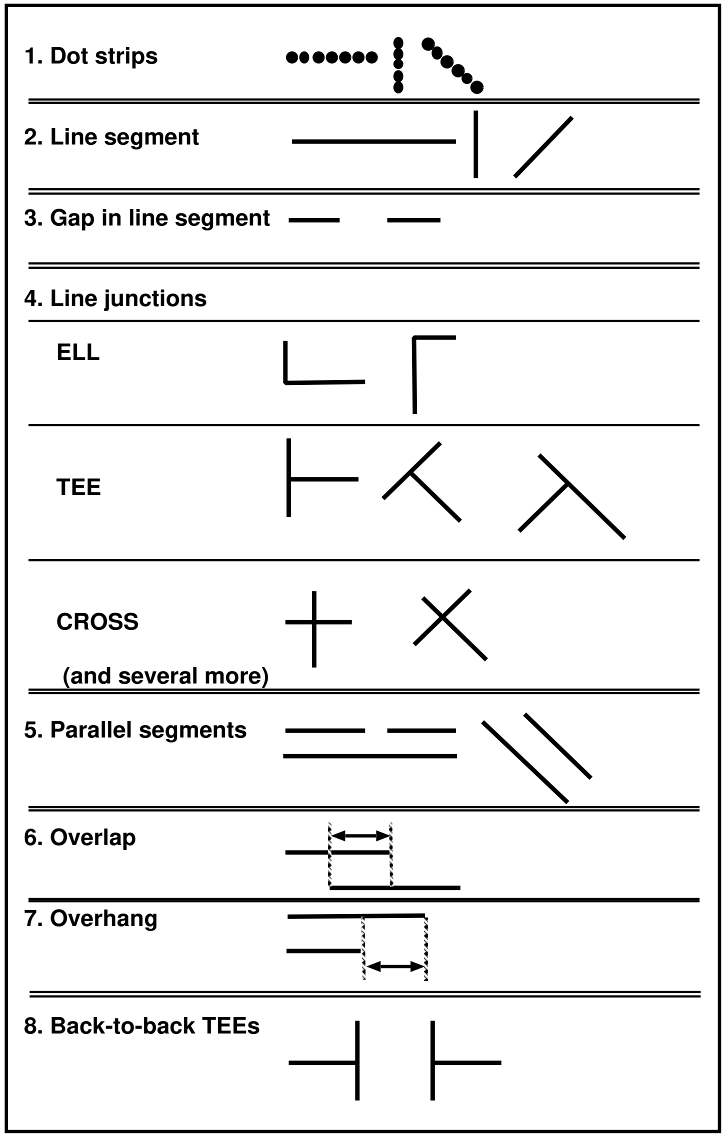
We seem to make use of structures of different sorts,

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

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

Only high level outputs are accessible to consciousness.

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



Putting it all together

We conjecture that a human visual system does concurrent processing at all these different levels of abstraction — and more.

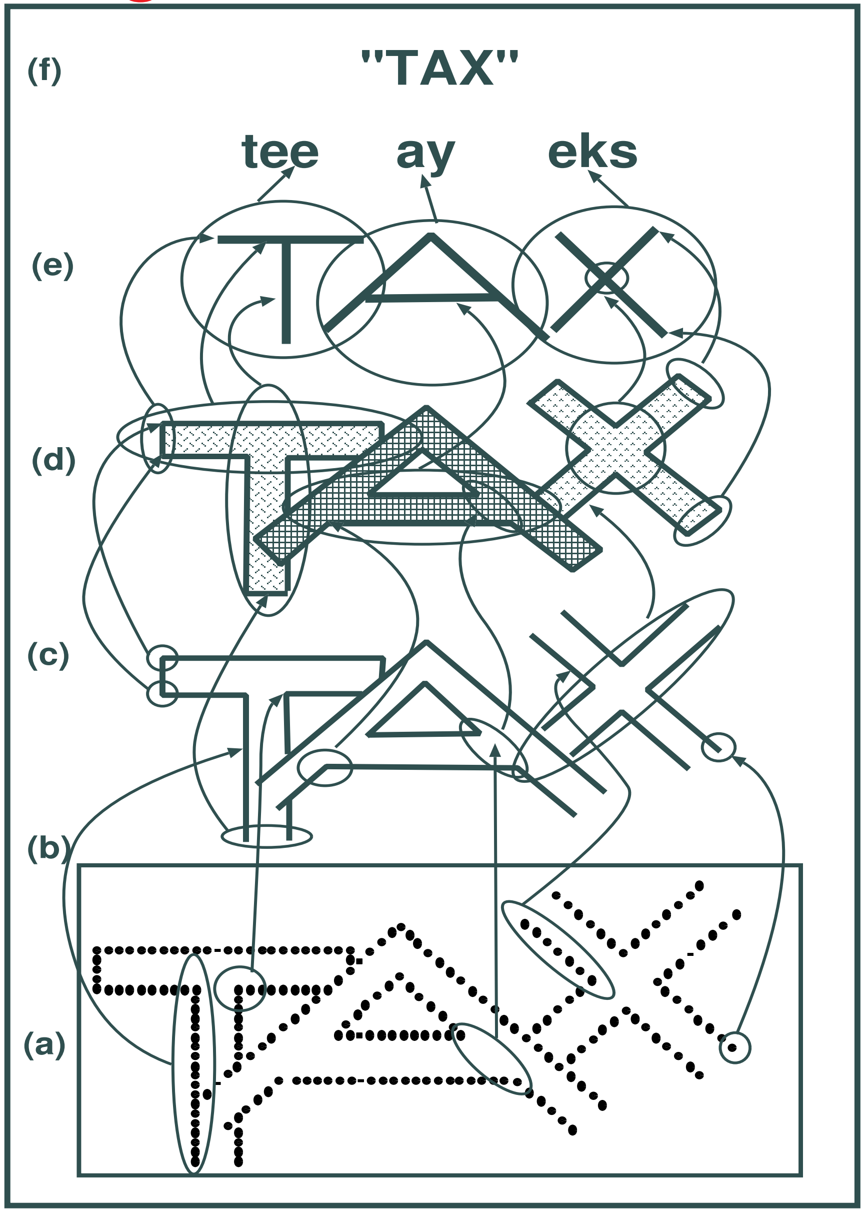
The POPEYE system (described in my 1978 book) was a first draft implementation.

Sub-systems at different levels could interact with other sub-systems, including interrupting them by providing relevant new information or redirecting “attention” or altering thresholds.

Sometimes a higher level subsystem (e.g. word recogniser) would reach a decision before lower levels had finished processing.

Sometimes the decision was wrong!

Humans also make visual mistakes.



Perhaps a network of neural nets could learn such things?

Conjecture: Animal visual systems use multi-layered architectures, with concurrent top-down and bottom-up flow of information, with multiple connections to central systems and to action systems (e.g. posture control, flight control).

Seeing affordances

Perception goes beyond seeing spatial structures, even beyond seeing structures at different levels of abstraction.

We also need to be able to see "affordances".

The psychologist J.J.Gibson introduced the idea of perceptual systems as detecting positive and negative affordances, e.g.

- graspability
- obstruction
- passage
- support
- rigidity
- a hook's ability to lift things

These affordances are concerned not just with what exists in the scene, but also what does not exist but might exist (e.g. grasping), or what could not exist (e.g. walking through a very narrow gap).
Some abilities to see affordances are the result of learning, others innate.

We don't know what collection of affordances a young child, an adult, a crow, a squirrel, an eagle, a sheep, a bumble bee can see.

Nor how they are represented in brains.

Using affordances

A purely reactive system like our simulated sheepdog implicitly detects affordances insofar as what it perceives immediately triggers appropriate actions (internal and external).

A deliberative system, like a nest-building crow, perhaps, or a typical human being, has to be able not only to perceive and react to affordances, but also to be able to represent alternative possible actions and the new affordances that those actions can generate.

This hypothetical reasoning about what might be or what might have been the case is most sophisticated in human beings, but it seems that some other animals are capable of it, to varying degrees. (See Kohler: *The mentality of apes*.) (Humans differ widely in their deliberative capabilities.)

The set of affordances available to humans in a typical cluttered spatial environment is huge.

Being able to detect them and select relevant ones and being able to learn about new affordances, are among the great achievements of ordinary human minds

IN ALTRICIAL SPECIES THE CAPABILITIES ARE NOT ALL THERE AT BIRTH.

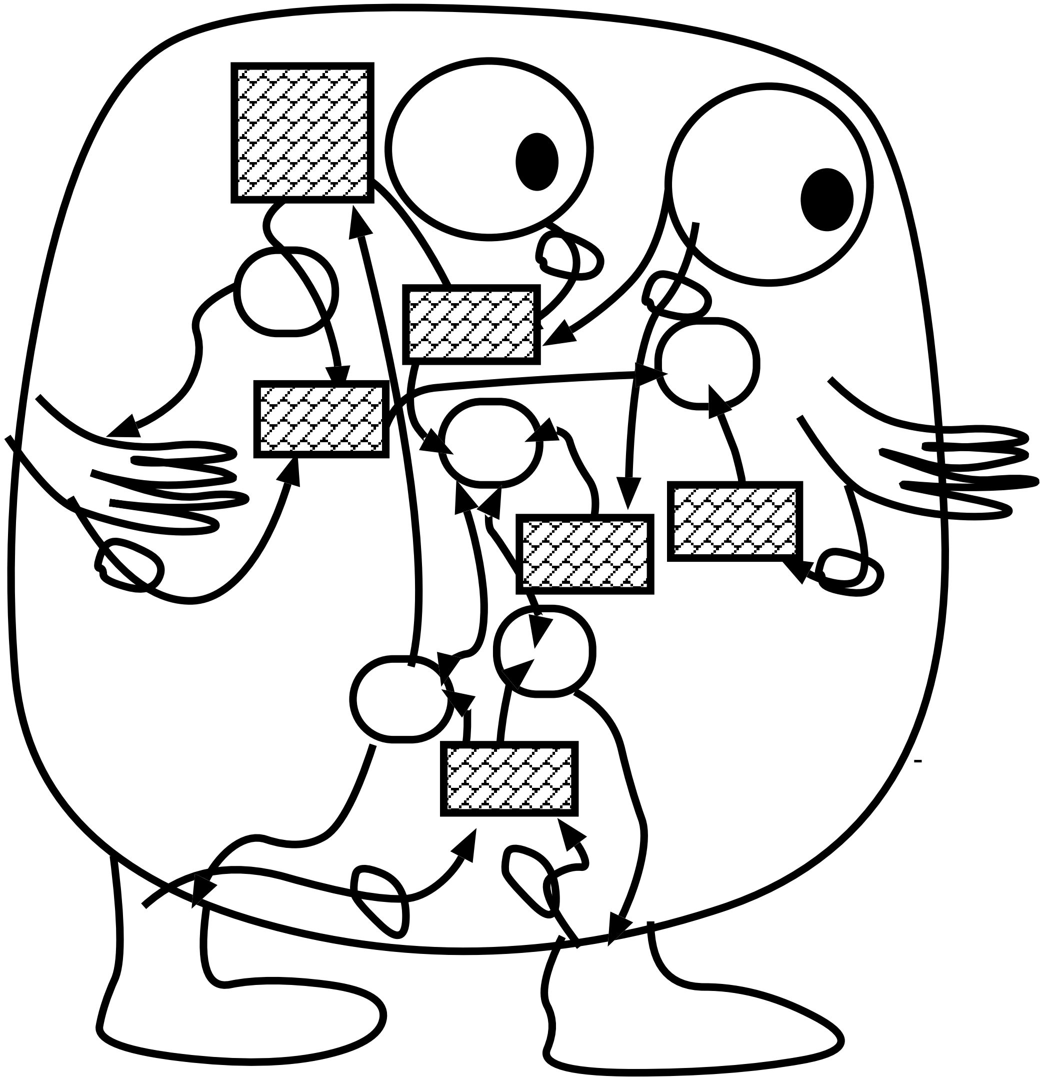
For a discussion of the innate/learnt trade-off, see this critique of “symbol grounding” theory: http://www.cs.bham.ac.uk/research/cogaff/talks/#talk14

How is all this implemented?

Could the architecture be an unintelligible mess?

Some researchers argue that we cannot hope to understand products of millions of years of evolution.

They work, but they do not necessarily have a modular structure or functional decomposition that we can hope to understand, or replicate.



Some of those people claim that the only way to produce machines with intelligence comparable to animals is to evolve them, e.g. using evolutionary computation methods, though that will not necessarily help us understand how they work.

Could the architecture be an unintelligible mess?

YES, IN PRINCIPLE

BUT:

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

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

Problem 2: the difficulty of changing a non-modular system without damaging old capabilities: evolution makes many thousands of changes over time.

Problem 3: storage space required to encode all possibly relevant behaviours if there's no "run-time synthesis" module.

CONJECTURE 1: evolution, like good engineers, 'discovered' the virtue of re-usable modules and and nearly decomposable complexes (H.A.Simon 1967).

CONJECTURE 2: evolution also 'discovered' that where there are separate modules they can usefully be combined "at run time" in different ways in different contexts. Hence the evolution of *deliberative* mechanisms.

Argued by Kenneth Craik, Karl Popper and many others.

The proposals below are presented within the framework of those conjectures.

Towards a unifying theory of architectures

- We need good general-purpose concepts for describing and comparing different classes of architectures for organisms and robots, and possibly other things.
- We build up our concepts by relating them to a space of possible architectures for integrated (non-distributed) agents.
- This space is characterised by a generic schema (a sort of grammar) specifying types of components and ways in which they may be related.

In the following slides we present a schema called CogAff. It is only a tentative first draft and will certainly have to be revised and enriched.

We do this by

– presenting different perspectives for dividing up an architecture

– showing how to overlay those perspectives to get a deeper understanding of the diversity

– indicating in a sketchy way how various aspects of human minds and other information processing systems relate to the various divisions in the architecture.

NOTE:
CogAff does not cover multi-agent architectures except insofar as the components of a single integrated architecture can be viewed as agents.

Perspectives on complete agents – 1.

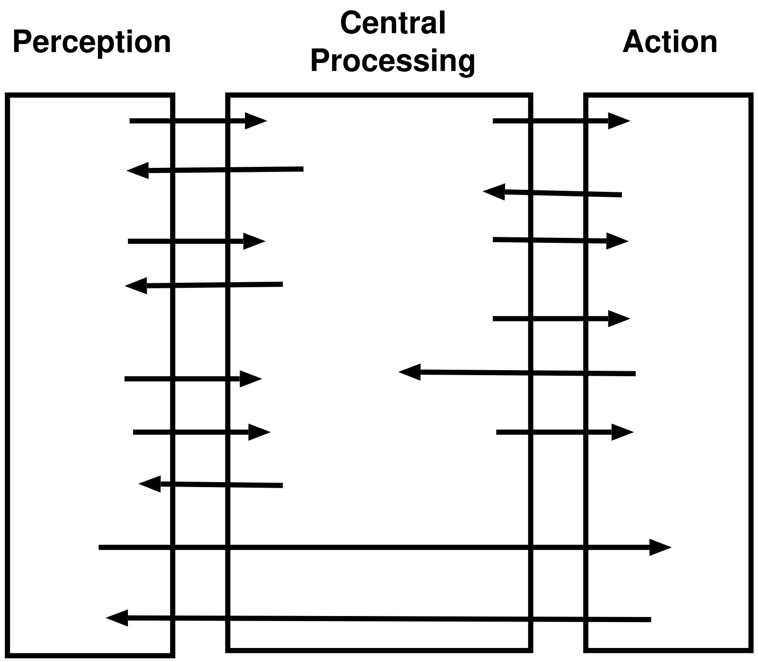
1. The "triple tower" perspective

(Many variants – Nilsson, Albus, ...)

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

Boundaries can change with learning and development.

What is the basis for distinguishing central from perceptual and action mechanisms: causal influence, and structural links?



A possible answer:

- The contents of the perceptual tower are largely under control of input from sensory transducers. Their function is primarily to analyse and interpret incoming information. They may also be 'in registration' with collections of sensory transducers.
- Similar criteria can be used for specifying contents of action tower.
- Contents of 'central' tower (a) change on different time-scales from those of perceptual and motor towers (b) are not closely coordinated with them.

A less obvious perspective – 2.

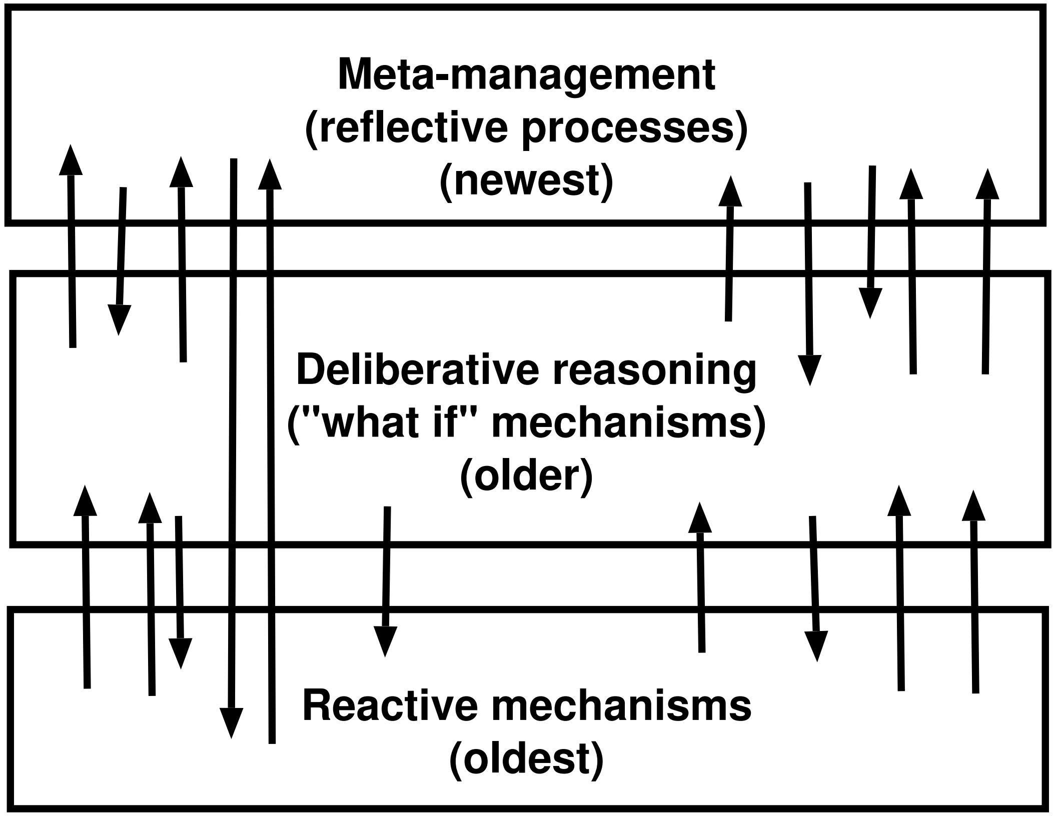
2. THE "TRIPLE LAYER" PERSPECTIVE

Another common architectural partition (functional, evolutionary).

There are many variants – for each layer.

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

Some people put reflexes in a bottom layer, seperate from more complex reactive mechanisms which include internal state changes.



Reactive mechanisms

- They are very diverse and may include many concurrent sub-systems, including "alarm" mechanisms.
- They may include forms of "parameter-adjustment" learning or adaptation.
- Some reflexes (innate or learnt) connect sensors directly to motors.
- Reactive mechanisms can be highly parallel and very fast.
- They may use analog or digital components or both.
- They may include neural nets, condition-action rule systems, lookup tables, decision nets, and other mechanisms.
- Some reactions change only internal state, affecting future reactions.
- Some internal states in a reactive system may act as implicit goals.
- It is difficult or impossible to program some reactive system directly or provide explicit information for them to use (compare neural weights.)
- But other parts of the system, or the environment, may "train" them.
- They make possible 'alarm-driven' primary emotions.
- They have been extremely successful in evolution: most individual organisms and most species are purely reactive.
  Most of the biomass on earth consists of reactive organisms,

So — why isn't everything purely reactive?

What can't reactive systems do?

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

However, for a purely reactive system to satisfy the same set of true counterfactual conditionals as a human deliberative system, it may require

- an impossibly long and varied process of evolution and training, in order to produce it

  and

- an impossibly large information store.

In contrast, 'a deliberative' system can combine primitive capabilities in new ways as needed in order to construct solutions to new problems – solutions not provided by training or by evolution.

Deliberative mechanisms

- Can represent and reason about non-existent or future possible entities.
- Some can also reason about what might have been the case in the past, or what may be the case out of sight.
- They allow alternative options to be constructed, evaluated, compared, selected then used.
- They can vary in the representational forms they use and the sophistication of their semantics.
  - Simple deliberative mechanisms may use only one step lookahead, and very simple selection mechanisms.
  - More sophisticated versions use compositional semantics in an internal language whose grammar admits unbounded complexity.
- They require a re-usable general-purpose working memory (garbage collectable).
- They require stored generalisations about what actions are possible in particular situations, and about the possible consequences of actions.
- They may be able to learn (new formalisms, new ontologies, new associations, ...)
- They benefit from perceptual systems that produce high-level chunked descriptions of the environment
- They may be able to train reactive systems that cannot be directly modified.
- They are typically slow, serial, resource limited (Why?) May need attention filters.
- They make possible secondary emotions using global 'alarm' mechanisms linked to deliberative mechanisms.

Natural deliberative systems are very expensive

Although a general-purpose computer can be used equally well to implement reactive or deliberative systems, it seems that in brains the mechanisms required for deliberative capabilities are very expensive, and therefore very rare.

- They require large associative information stores with sophisticated learning capabilities
- They depend on elaborate perceptual mechanisms that extract reusable information from sensory data at a useful level of abstraction
- They require forms of information encoding that may not map naturally onto biological mechanisms, and therefore require support for appropriate virtual machines
- They need a kind of general-purpose re-usable memory within which it is possible to create temporary descriptions with variable structures, including descriptions of descriptions.
  (Garbage collection)

Unanswered questions about intermediate cases.

- We understand how to design a fairly wide class of purely reactive systems.
- We understand how to design certain deliberative systems (though most of them are limited by problems of controlling search).
- We don't know enough about possible intermediate cases:
  - Proto-deliberative systems are reactive systems in which alternative responses can be triggered simultaneously, requiring some form of *arbitration mechanism* to resolve conflicts.
    (Some people label these "deliberative".)
  - There may be many varieties of formalisms of varying complexity and power in intermediate mechanisms.
  - There may be different sorts of re-usable short term memories for constructing options to be evaluated, with different limitations.
    E.g. some have more restrictive structural depth limits than others.
  - There may be different modes of comparison and evaluation of alternatives
    (e.g. compare weights vs difference descriptions)
  - There may be different forms of storage of information about possible actions in a context, or possible extensions of an incomplete structure.

These and other ideas may be relevant to thinking about evolutionary and developmental trajectories.

Meta-management mechanisms

- They can monitor, categorise, evaluate, and (to some extent) control other internal processes – e.g. some deliberative processes, or some perceptual processes. (See Barkleys's 1997 book on ADHD)
- This includes control of attention, control of thought processes. (Such control is lost in tertiary emotions.)
- Both monitoring and control depend on special purpose low level support for new architectural information pathways.
- They can vary widely in sophistication, e.g. depending on social learning.
- They require concepts and formalisms suited to self-description, self-evaluation
- They support a form of internal perception which, like all perception, may be incomplete or inaccurate, though generally adequate for their functional role.
- The concepts and formalisms may be usable in characterising the mental states of others also.
- Different meta-management control regimes may be learnt for different contexts (different socially determined “personae”).
- Evolution of sensory qualia: occurs when it is useful for meta-management to look inside intermediate levels of perceptual processing (why?).
- If meta-management mechanisms are damaged, blind-sight phenomena may occur. (Experiments requiring subjects to *report* what they see typically use the meta-management layer! What's happening in other layers may be unnoticed.)

Varieties of meta-management

- Different types of meta-management can use more or less sophisticated forms of representation and processing.
- They can also vary in the types of evaluation they can apply
- In humans much self-categorisation and self-evaluation is socially/culturally determined. (E.g. feelings of guilt or sin)
- The existence of meta-management may provide a “niche” encouraging evolution of higher level *perceptual* mechanisms categorising mental states of other agents. (Top-left box in grid diagram. Likewise top-right box for action mechanisms.)
- This may have required parallel evolution of involuntary “expressive” behaviours (See: Sloman 1992 on the dangers of complete voluntary control of sincerity.)
- The absence of meta-management was a major factor in the fragility and incompetence of many old AI systems (e.g. they could not tell when they were reasoning in circles, or solving a minor variant of a previously solved problem.)
- Mechanisms for triggering and modulating meta-management processes may produce a far wider variety of short term and long term affective states than scientists have so far categorised. (Contrast novelists!)
  (E.g. see the analysis of grief in Wright, Sloman and Beaudoin (1996) at CogAff web site.)
- Some researchers (e.g. Damasio) appear to be confused about how the requirement for meta-management relates to requirements for emotional mechanisms, and wrongly infer that emotions are *required* for intelligence.

More on architectural layers

The layers differ in:

– Evolutionary age (reactive oldest).
– Level of abstraction of processing (reactive least abstract),
– The types of control functions, and mechanisms used
  (e.g. ability to search, evaluate, compare; amount of parallelism; use of neural vs "symbolic" mechanisms)
– The forms of representation used
  (e.g. flat vs hierarchical compositional syntax and semantics)
– The varieties of structural change they support

The distinctions between layers are not necessarily very sharp, and there can be intermediate cases.

In fact it is very likely that evolution produced intermediate cases.

But we don't yet have a clear view of the space of possible designs or the space of possible niches relevant to varieties of meta-management.

Note: talk of "layers" is just a first approximation: the different types of mechanisms and formalisms may not be always be grouped into distinct layers or sub-structures. Rather, the different kinds of formalisms and mechanisms may be found within various portions of the architecture.

(Compare Minsky's draft book: *The Emotion Machine* at his website.
He divides up the functionality discussed here in a different way.)

“Layered” architectures have many variants

Later we’ll see that designers present layered architectures with different subdivisions and different interpretations of subdivisions, and different patterns of control and information flow.

Divisions between layers can be based on:

- evolutionary stages
- levels of abstraction,
- kinds of formalism used
- control-hierarchy, (Top-down vs multi-directional control).

There can be different kinds of information flow between layers.

E.g. sometimes information flow is treated as a sort of “pipeline” (e.g. the popular ‘Omega’ Ω model of information flow, described below).

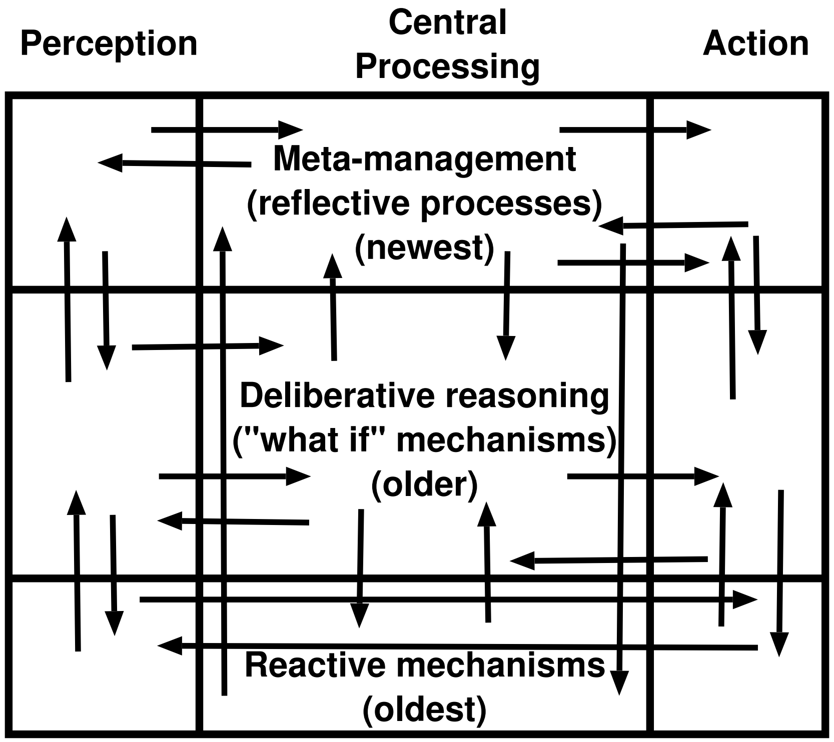
We have tried to to make the CogAff framework subsume many such design options.

LAYERS + PILLARS = CogAff GRID

We can overlay the two perspectives, giving a grid of co-evolved sub-organisms, each contributing to the niches of the others.

THIS IS AN ARCHITECTURAL "SCHEMA" SPECIFYING POSSIBLE COMPONENTS AND RELATIONSHIPS BETWEEN COMPONENTS, NOT AN ARCHITECTURE.

The CogAff schema defines a variety of components and linkages.
Not all the components, and not all the communication links, need be present in all species of natural or artificial architecture.



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

Information may flow in ways not shown by the arrows - e.g. diagonally across layer boundaries. (Example?)

This is a very general schema, whose instances can differ greatly. Contrast the H-Cogaff (human) instance (below).

The “Omega” model of information flow

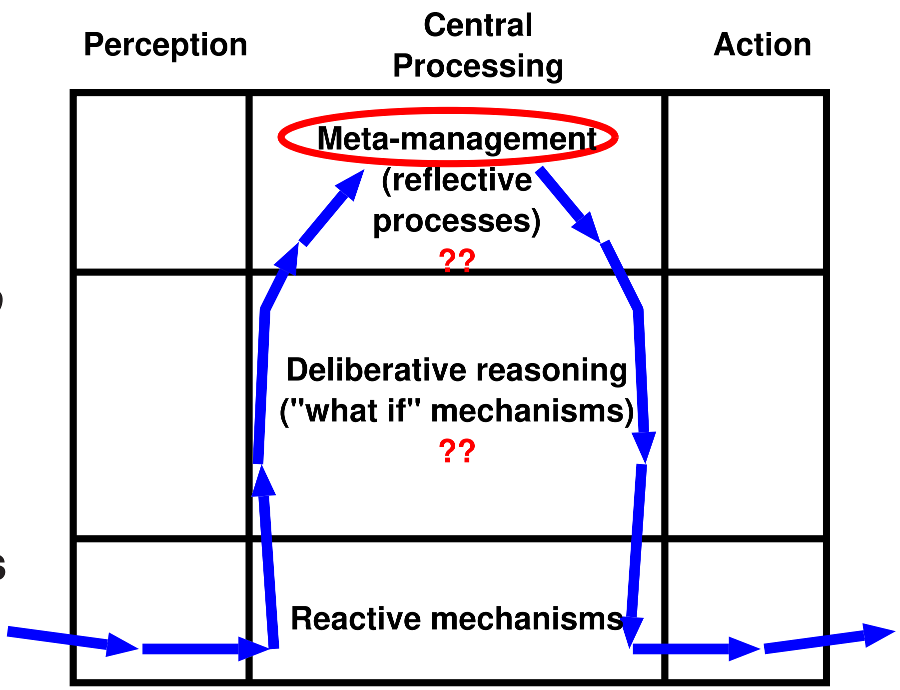
CogAff allows many variants including ‘Omega’ architectures.

E.g. the “contention scheduling” model (Cooper and Shallice 2000).

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

Shallice has proposed a “Supervisory Attentional System” (SAS).

Other suggestions are found in Barkley’s book on ADHD, and in much speculation about “executive functions” by psychologists and neuroscientists.



The “Omega” model does not allow for layered concurrent perceptual and action towers separate from central tower.

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

TENTATIVE ANSWER:

Multi-level (multi-window) perception uses dedicated concurrent parsing and interpretation of sensory arrays, e.g. building new data-structures in registration with sensory arrays, e.g. in registration with a 2-D visual array.

Contrast “peephole” perception.

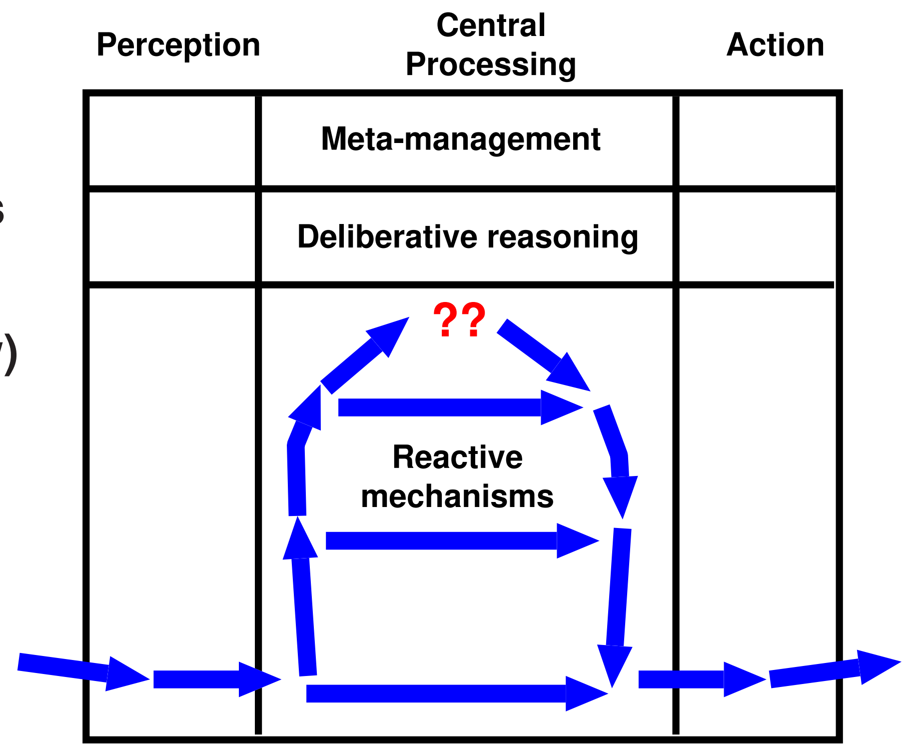
Likewise we can distinguish “multi-window” and “peephole” action architectures.

Another special case of CogAff: Subsumption architectures (Brooks)

These allow different architectural layers, but only within the reactive sub-space, where they form a sort of dominance hierarchy (unlike the layers in H-Cogaff described later.)

Some Brooksians (but not Brooks now) deny that animals (even humans) use deliberative mechanisms.

(How do they get to overseas conferences?)



These reactive subsumption architectures are able to meet requirements for human-like capabilities ONLY IF quite unrealistic assumptions are made about evolutionary developments, storage capabilities, etc.

Different architectures use different subsets of CogAff

Subsumption, uses only a subset of the mechanisms allowed in the CogAff schema.

The Omega architecture uses a different subset.

Some AI systems use only deliberative mechanisms.

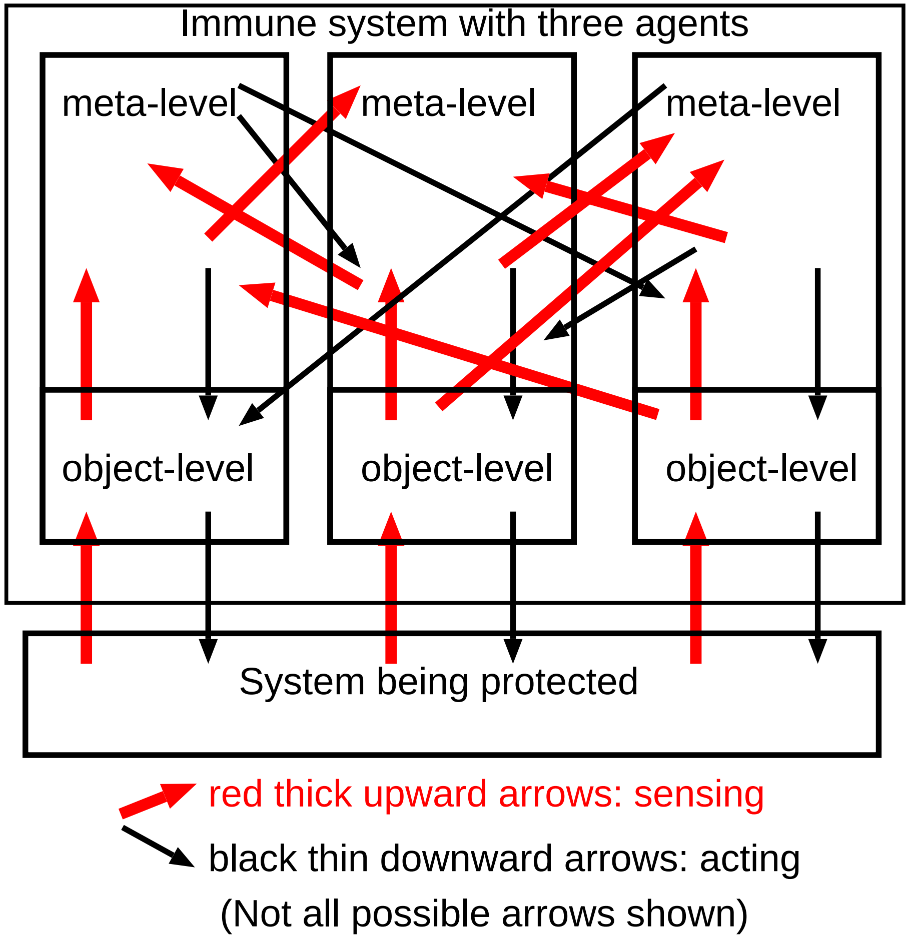
> We should avoid all dogmatism and ideology, and investigate which subsets are useful for which organisms or machines, and how they might have evolved.
>
> That way we'll learn instead of fighting.

A mutual meta-management system

Catriona Kennedy has extended these ideas in the design of a system for detecting and repairing code damaged by hostile intruders.

To avoid the fragility of having only one monitor, Kennedy proposes a collection of them each observing not only the system being protected but also one another's observations, and, if appropriate, taking "corrective" action, e.g. repairing damaged code.

The "object level" components monitor and act on the system being protected. The meta-level components monitor and act on the object- and meta-level components (which may be reactive, deliberative or a mixture).



Some of Kennedy's papers outlining the theoretical ideas and describing a prototype implementation can be found here: http://www.cs.bham.ac.uk/research/cogaff/0-INDEX00-05.html

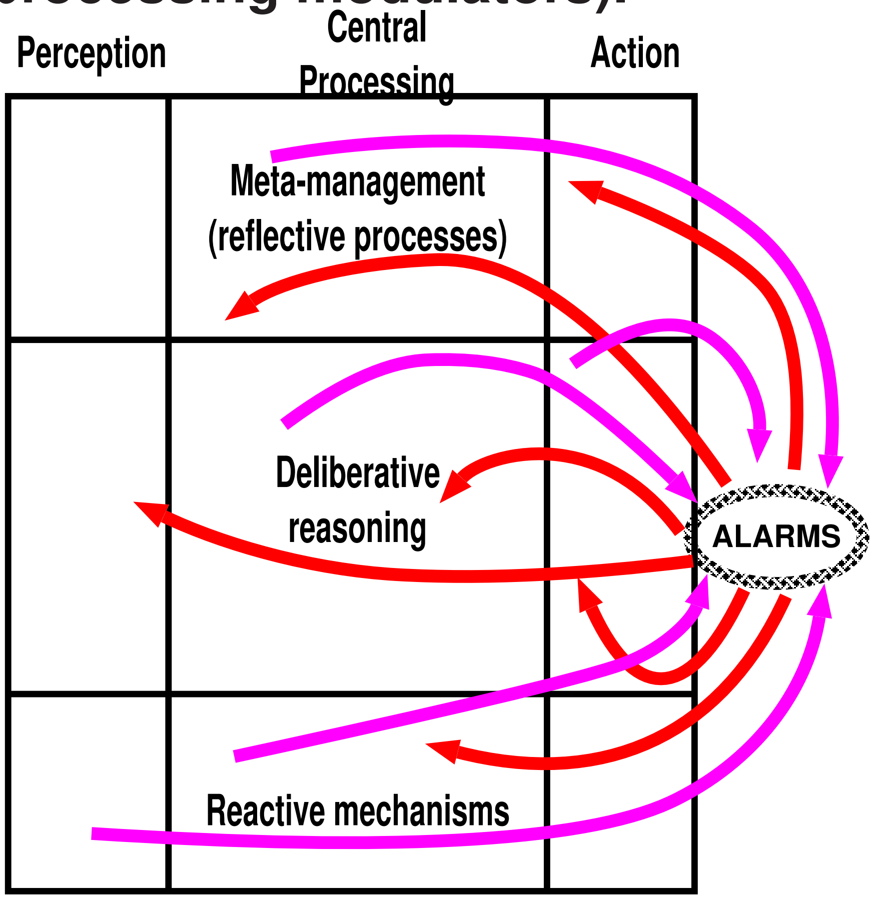
The need for “alarm” mechanisms

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

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

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

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



The five Fs: Feeding, fighting, fleeing, freezing, and reproduction
(Usually only four are specified!)

Many sorts of alarms

- **Alarms allow rapid redirection of the whole system or specific parts of the system required for a particular task (e.g. blinking to protect eyes.)**
- **The alarms can include specialised learnt responses: switching modes of thinking after noticing a potential problem.**
- **E.g. doing mathematics, you suddenly notice a new opportunity and switch direction. Maybe this uses an evolved version of a very old alarm mechanism.**
- **The need for (POSSIBLY RAPID) pattern-directed re-direction by meta-management is often confused with the need for emotions e.g. by Damasio, et. al.**
- **Towards a science of affect:**

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

Self-aware information-processing machines

Humans have not only *reactive* and *deliberative* abilities but also meta-management abilities.

They have the ability to

- **Monitor and categorise some of their internal states and processes (e.g. thoughts, reasoning strategies, desires, percepts).**
- **To evaluate those internal states as good or bad**
- **To control or modify some of them (but not all)**
- **To use the same concepts in perceiving and thinking about other humans, animals — and maybe robots.**
  (E.g. seeing someone as happy, sad, interested.)

These high level self-monitoring and self-managing capabilities are limited.

- **We can't monitor everything going on inside us (e.g. you can't introspect to find out the grammatical rules you know).**
- **We can't control everything we try to control, e.g. we sometimes can't stop our thoughts wandering off the current task.**
  Certain kinds of emotion involve partial loss of control of attention.

Summary so far

- All animals, including humans, have some reactive mechanisms.
  Some animals (e.g. insects?) don't have anything else.
- We can expect robots moving about in a physical world to need reactive mechanisms.
- Some animals (perhaps more than we realise) also have deliberative capabilities, though they vary in the depth and variety of those capabilities.
- Human-like robots will also need to be able to think about what might happen if, and to consider and compare possible future actions.
- At least humans, and perhaps also a few other species have meta-management capabilities that involve internal observation, evaluation, and, to some extent, control. The descriptive categories required for this can also be used in perceiving and thinking about what's going on in other agents (e.g. ones that might eat you or be good to eat).
- Human-like robots will also require these reflective, meta-management capabilities and the corresponding extended perceptual abilities.

There is still much we don't know about how to make these requirements for human-like robots sufficiently precise to be the basis of working designs. There is still much work to be done.

See also Marvin Minsky's draft book *The Emotion Machine* at his web site:
http://www.media.mit.edu/˜minsky/

The architecture of a human mind

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

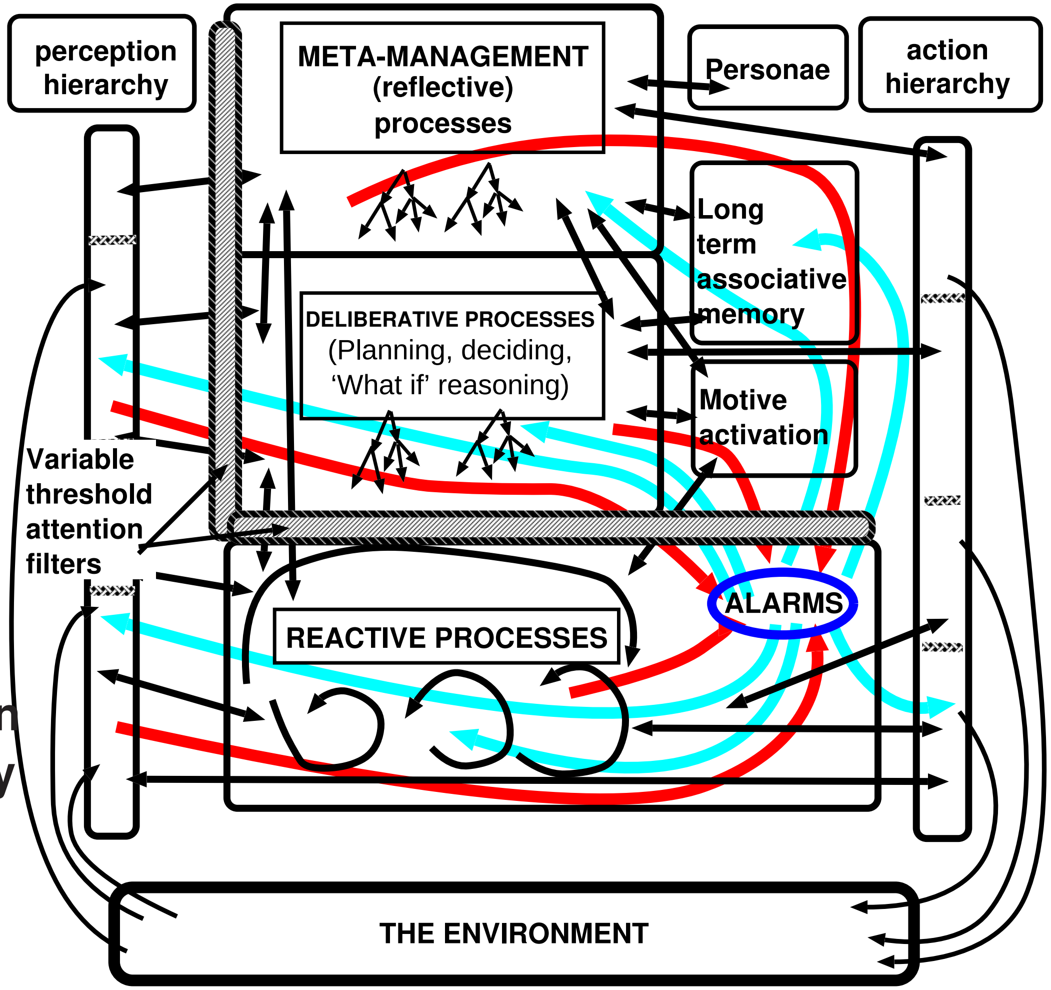
This conjectured architecture (H-Cogaff) is a particularly rich special case of CogAff.

It could be included in human-like robots (in the distant future).

Arrows represent information flow (including control signals)

‘Alarm’ mechanisms can achieve rapid global re-organisation.

If meta-management processes have access to intermediate perceptual databases, then this can produce self-monitoring of sensory contents, leading philosopher robots with this architecture to discover “problem(s) of Qualia?”



This has profound implications for education: how do the different components develop and how can we help or hinder that process?

Implications of the architecture

Within the framework of an architecture like this we can explain

- Many different forms of perception, involving different levels in the perceptual mechanisms.
- Different kinds of action control, involving different levels in the action mechanisms and different central processes.
- Different kinds of learning and development: including changes in sub-systems, and changes in connections between sub-systems (e.g. training of the reactive sub-system by actions controlled initially by the deliberative mechanisms).
- Different kinds of emotions: including effects of the alarm mechanism provoked by events in different parts of the system and causing changes in different parts of the system.
- Other affective states, including desires, moods, long term goals, attitudes, etc.
- Limitations of humans and other animals, e.g. because the deliberative mechanisms are inherently slow and serial and resource limited because they are expensive (at least in brains).

For more on all this see the CogAff papers:

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

Alternative architectures

Not all animals, and not even all humans (e.g. newborn infants) have the full variety of architectural components and linkages indicated in the previous slide – the H-CogAff architecture.

The architecture inspires many research questions, e.g.

- **Which components in the architectures are present in which animals?**
- **How did such an architecture evolve, and what biological requirements led to the evolution of different parts of the architecture.**
- **How can the architecture grow or develop: what sorts of changes occur from infancy onwards in a human mind?**
- **How many different kinds of learning occur in an architecture like H-CogAff**
- **Can understanding this help us develop new educational strategies, instead of the hit-and-miss swings of fashion in education that we have seen so often in the past.**

Perhaps if we teach children to design and test and debug and modify and explain working examples of such architectures, they will grow up into scientists capable of developing deep new ways of thinking about human minds and how they work?

There are many other AI research areas

I have glossed over many questions investigated by AI.

How does human language work?

- ***How do we understand sentences we have never heard before?***
- ***How do we produce sentences we have never heard before?***
- ***How can we think about and talk about things that do not exist (e.g. Father Christmas, Harry Potter, Darth Vader, heaven?)***

How do we do mathematical reasoning

- ***including thinking about infinite sets (like the set 2, 4, 6, 8, 10, 12, ...)?***

How do we select and control our actions?

How do we make plans

How do we learn (concepts, theories, skills, languages, ...)

Some of these questions arise for other animals also.

And for AI as engineering they arise for robots and intelligent software systems.

Some requirements for progress

- We must not assume that we know what the capabilities of humans and other animals are: we have to continue studying them as engineers.
- We must not assume that we know what sorts of information processing systems are possible, or that we have a good understanding of possible ways of programming computers.
- We shall have to explore novel information-processing architectures
- We shall have to explore novel forms in which information can be stored

  (not just familiar ones: numbers, equations, logical formulas, bit-patterns, arrays, tree-structures, graph-structures, neural nets.)
- We shall have to develop novel mechanisms for creating, manipulating, and using these information-bearing structures.
- We shall probably have to develop new kinds of mathematics in order to formalise and analyse what we are doing and discover its limitations.

AI is inherently multi-disciplinary

Many of the questions are also studied, though in different ways, by other disciplines. E.g.

- **Psychology and brain science** study perception, learning, language use, reasoning.
- **Linguistics** studies the detailed structures of different languages, and how they developed.
- **Philosophy** investigates the nature of thought and language and reasoning, the relations between mind and body, the nature of science, and much else.
- **AI needs to interact with these other disciplines in order to benefit from their theories and in order to take account of the full range of phenomena to be explained.**
- E.g. from psychology and brain science we can learn some of the very strange things that can happen if a bit of your brain is damaged: you may have some *parts* of a previous ability still working while other parts don't work.
- Unfortunately the structure of our educational system and the tremendous pressures on university staff (publish or perish) make it very difficult nowadays for people to acquire a broad and deep multi-disciplinary education.

RADICAL REFORMS IN OUR THINKING ABOUT TEACHING AND RESEARCH AT ALL LEVELS WILL BE NEEDED, TO PRODUCE SCIENTISTS OF THE RIGHT CALIBRE.

Brains support consciousness? How?

What's consciousness?

People assume consciousness is one thing.

Then they ask questions like:

- **which animals have IT?**
- **how did IT evolve?**
- **what is ITS function?**
- **could machines have IT?**
- **which bits of the brain produce IT?**

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

- What we call "consciousness" is a large ill-defined COLLECTION of capabilities.
- THEY (the various capabilities) can be present or absent in different combinations, in different animals, in people at different stages of development or after brain damage.
  Also in different machines.
- No pre-ordained subset of that set of capabilities is THE subset required for consciousness.
  - Compare flea, fish and frog consciousness.
  - Compare infant and adult human consciousness.
- Not just ONE thing that is always present or absent. Neither is it a matter of degree.
- I.E. "CONSCIOUSNESS" IS A PARTLY INDETERMINATE "CLUSTER CONCEPT".
  (Like "emotion")
- People think they know what IT is from experience.
  Before Einstein people thought they knew what simultaneity was from experience. We can unintentionally fool ourselves.

(The notion of a "cluster concept" is explained briefly in this slide presentation:
http://www.cs.bham.ac.uk/research/cogaff/ibm02/)

Varieties of consciousness

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

- **Microbe consciousness**
- **Flea consciousness**
- **Frog consciousness**
- **Eagle consciousness**
- **Chimp consciousness**
- **Infant (human) consciousness**
- **Adult consciousness**
- **Varieties of drug-modified consciousness**

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

Towards a conclusion ....

Understanding what human beings are, and being able to design an implementable human-like robot requires us, at least, to understand:

– the varieties of affordances in human environments, and ways in which affordances can be perceived, represented, and used in acting in the environment.

– the varieties of types of co-evolved concurrently active information-processing sub-systems that make up a human being: the CogAff schema provides only a first draft, coarse-grained, taxonomy.

– how all these subsystems develop, and how they interact with one another, and what sorts of states they can generate (e.g. varieties of emotions, moods, pleasures, pains, and other affective states).

– the varieties of forms of representation deployed within the different subsystems, and how they are used in learning various ontologies that develop during a person's life.

– the varieties of ways in which the system can go wrong, producing both genetic malfunctions and manifestations of various kinds of brain damage and disease. (E.g. R.Barkley on ADHD)

– the capabilities that appear to exist only in humans, and why they do not occur in other animals. This includes understanding how much of a typical human mind is genetically programmed, how much a social product, etc.

Different information-processing techniques may be required for complete human-like systems.

including:

- Symbol-manipulating programs
- Logical reasoning systems
- Neural nets
- Genetic algorithms (evolutionary computation)
- Dynamical systems (based on models from physics)
- Where appropriate, new sorts of hardware

Do not believe anyone who tells you that only one kind of technique works: some ignorant people have narrow-minded views of AI

To support the development of these and other techniques, AI researchers have designed especially powerful and flexible programming languages, e.g. Lisp, Prolog, Pop-11, Scheme, Rule-based languages, constraint languages, ...

For more on this see: http://www.cs.bham.ac.uk/~axs/misc/talks/#talk11

Using other languages for AI research, e.g. C++, Java, is possible, but typically it slows down development because these languages are not so suitable for incremental, exploratory development of ideas about very complex problems.

SUB-FIELDS OF AI

AI has many sub-areas studying different problems.

Examples are

- **Natural language processing**
- **Vision**
- **Learning**
- **Automated discovery (induction)**
- **Memory**
- **Problem solving**
- **Theorem proving**
- **Planning**
- **Modelling motivation and emotions** **(e.g. in entertainments)**
- **Expert systems**
- **Robotics**
- **Motivation and emotions**
- **Architectures for integrated multi-functional minds**
- **Evolution of intelligence**
- **(... and many more ...)**

Unfortunately, too often the people working in one subfield do not think about how the results of their work should be combined with results from other subfields in complete working systems.

And finally

We are only now beginning to understand
what information-processing machines are
what intelligence is
what we are.

WHY NOT JOIN IN THE ADVENTURE?

Some sources of information

http://www.cs.bham.ac.uk/~axs/misc/aiforschools.html
http://www.aaai.org/aitopics/
(There are also many textbooks on AI referred to at those web sites.)

Online slide presentations on this and related topics are at this location:
http://www.cs.bham.ac.uk/~axs/misc/talks/