EC Cognitive Systems ‘Kickoff’ Conference

Tutorial on Integration

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

School of Computer Science
The University of Birmingham, UK

I shall put these slides on the Birmingham CoSy web-site

http://www.cs.bham.ac.uk/research/projects/cosy/

Slides produced and presented using Linux only

(Slides still under development)

Integration in Cognitive Systems

The EC Cognitive Systems Initiative addresses several kinds of integration.

ftp://ftp.cordis.lu/pub/ist/docs/cs-support-document-v2.pdf

I shall discuss only a subset concerned with re-integrating the study of intelligent systems. This research has become increasingly fragmented in the last three decades.

Integration in Cognitive Systems

The EC Cognitive Systems Initiative addresses several kinds of integration.

ftp://ftp.cordis.lu/pub/ist/docs/cs-support-document-v2.pdf

I shall discuss only a subset concerned with re-integrating the study of intelligent systems. This research has become increasingly fragmented in the last three decades.

The fragmentation includes

- Problem areas (e.g. vision, language, learning, reasoning, ...)
- Disciplines (e.g. AI, Linguistics, Psychology, Biology...)
- Concern with humans vs other animals
- Focus mainly on adult human competence, ignoring development
- Disagreements on philosophical positions:
  - GOFAI,
  - Symbol grounding.
  - Behaviour-based AI,
  - Dynamical systems,
  - Embodiment (and situatedness?,)
  - Logic vs other forms of representation

Mostly silly debates: people argue as if some useful point were the whole truth.
Wheels get re-invented, possibly with new names.

We need more backward-chaining research

Most researchers most of the time start from where they are and ask

> **What can I do next to extend, improve, generalise my work?**

Contrast research driven by the question:

> **What don’t we know that we should aim to know in 20 years time, or 50 years time, and what do we need to do now to move in that direction?**

The first is ‘forward chaining’ research.

The second is ‘backward chaining’ research.

We need more backward-chaining research

Most researchers most of the time start from where they are and ask

What can I do next to extend, improve, generalise my work?

Contrast research driven by the question:

What don’t we know that we should aim to know in 20 years time, or 50 years time, and what do we need to do now to move in that direction?

The first is ‘forward chaining’ research.

The second is ‘backward chaining’ research.

- Forward chaining research often leads to increasing fragmentation, narrowness of vision, re-invention of history, demonising other approaches and failing to notice some hard, important problems.
- But backward chaining research is far more risky and in the current research funding/assessment climate and conditions for tenure there are many institutional pressures against it.

Not everyone can do, or should do, backward chaining research, but if too few people do it, science suffers in the long term.

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

The UK Computing Research Grand Challenge proposal on ‘Architecture of Brain and Mind’ (GC5)

Some examples

Compare

- trying to make your ‘sentence-understanding system’ perform slightly better than someone else’s on a standard test corpus

with

- trying to combine your sentence understanding system with the ability to have a conversation about re-arranging furniture in a room.

Compare

- trying to make your face recognition program do better than someone else’s on a set of standard images

with

- using face perception in a conversation: judging mood, concern, focus of attention (what the speaker is looking at), agreement, puzzlement, boredom, friendship, eating with a tender tooth...

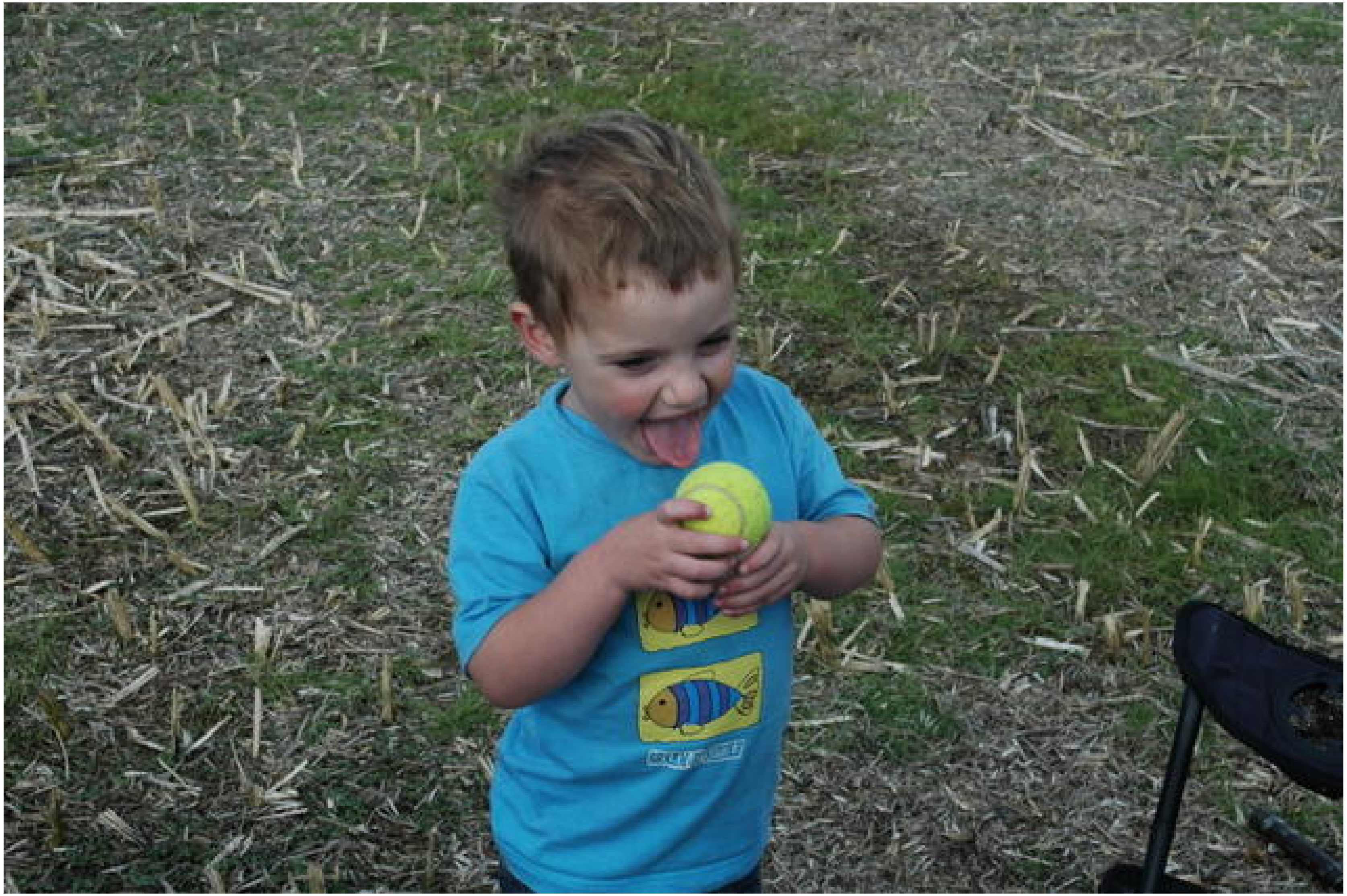
Compare

- making your program recognize a ball

with

- Seeing a ball held in a shuttlecock as an ice-cream

Seeing a visual metaphor



Putting the components together

From Colette Maloney's presentation on the EC Cognitive Systems initiative June 2003

... from low-level processing & robustness of individual components to a systems approach where all components - including high-level cognitive functionalities - have a role to play in assuring robustness ... working acros areas, combining biological vision, AI, computer vision ...

- **Focus is on:** methodologies and construction of robust and adaptive cognitive systems integrating perception, reasoning, representation and learning, that are capable of interpretation, physical interaction and communication in real-world environments for the purpose of performing goal-directed tasks.
- A main target of this research is interdisciplinarity, i.e., to carefully consider the integration of different disciplines including computer vision, natural language understanding, robotics, artificial intelligence, mathematics and cognitive neuroscience and its impact on overall system design.

Crucial question:

What does it mean to *integrate* perception, learning, high level cognition, biological vision, etc. etc. ?

What do we need to integrate?

Some of the answers are fairly obvious

- Integration of multiple functions
  (perception, action, language, learning, wondering, ...)
- Integration of forms of representation
  visual, logical, neural, symbolic, sub-symbolic,linguistic, procedural,...
- Integration across time and place
  episodic memory, generalisation, explanation, prediction, saccades, two hands, routes,...
- Integrating ordinary perceptual knowledge with deep theoretical science.
  ‘This is rigid’ vs ‘This has a crystalline structure’
- Integration of old and new mechanisms
  reactive subsystems, alarms, deliberative, meta-semantic, reflective ...
- Integration of cognition and affect
  wanting, preferring, liking, disliking, intentions, attitudes, moods, ...
- Virtual and physical machines: mind and body
  deciding to walk, contracting muscles, using gravity, compliant wrists ...
- Integration of individual intelligence with social processes
  nature/nurture, learning/development, absorbing a culture, fighting, bargaining
- Integration of tools and design ontologies used by different researchers who normally don’t work together.

But the details are not at all obvious

One approach to understanding the details:

- Collect many examples of human (or animal) competence
- Don't make the common mistake of ignoring precocial species

  deer, chickens, ... competent at birth/hatching

  — then what has evolution done for altricial species (lions, eagles, humans)?

  Compare: http://www.cs.bham.ac.uk/research/cogaff/talks/#meanings
- Study the examples, analyse and compare them them

  (often deceptively hard: things may be far more complex than they seem)
- Replicate
- Debug
- Repeat
- Look for new examples and extend...

Look at some real life examples to gain:

- new, grander, long term ambitions
- deep humility

SHOW DEMOS

- Show competent Crow
- Show incompetent Child

How these two examples should be analysed is far from clear.

There are conceptual as well as empirical questions

- Show simplified SHRDLU
- Show Deb Roy’s responsive Ripley

What sort of demo should impress us?

- Show dog
- Show child with train and tunnel

A child sitting on the floor, surrounded by a train set, is constantly seeing different things, yet clearly acts as if in a fixed (though changeable) environment most of which is known even when not being perceived.

Unnoticed ontological blindness

- Scientists get trapped by their own conceptual schemes unable to see the distortions and omissions in their description of what needs to be explained.
- Concepts get re-defined so as to fit what the theory says – e.g. defining ‘consciousness’ in terms of a particular mechanism, then saying “Look my system models consciousness”.
  Likewise perception, learning, reasoning.....

  See paper by Sloman&Chrisley in *Cognitive Systems Research* 2004.

Partial solution:

talk to

and listen to!

people from other approaches, not just your own in-group.

Unobvious complexities

A trap to avoid: replicating only restricted aspects of a competence, not noticing the missing bits:

- A visual system recognising only isolated instances, e.g. not dealing with cluttered scenes.
- Recognising instances of a category but not seeing their shape
- Recognising but not understanding relations between views of the object.
- Not using the full set of categorisations (e.g. ‘cow’, but not: ‘animal’, ‘farm animal’, ‘living thing’, ‘grazing’, ‘walking’, ‘looking nervous’, etc.....)
- Assuming that the ability to recognize and label instances of a class gives meaning to the labels used, ignoring the meaning that is determined by role of the concept in a theory of the world (ignoring theory-based meaning).

  See http://www.cs.bham.ac.uk/research/cogaff/talks/#meanings
- Recognising objects, but not seeing any actions, or motion of the object.
- Robots that perform a task, but don’t know what they haven’t done or why.
- Recognising, but not being able to do anything with or understand use of e.g. a tool, eating utensil, garden implement.
- Recognising objects but not seeing the space between or around them.

  HOW CAN EMPTY SPACE BE SEEN? – more on this later.

Integration WITHIN a function

Besides integration of different functions,

e.g. vision, language, walking, grasping, ...

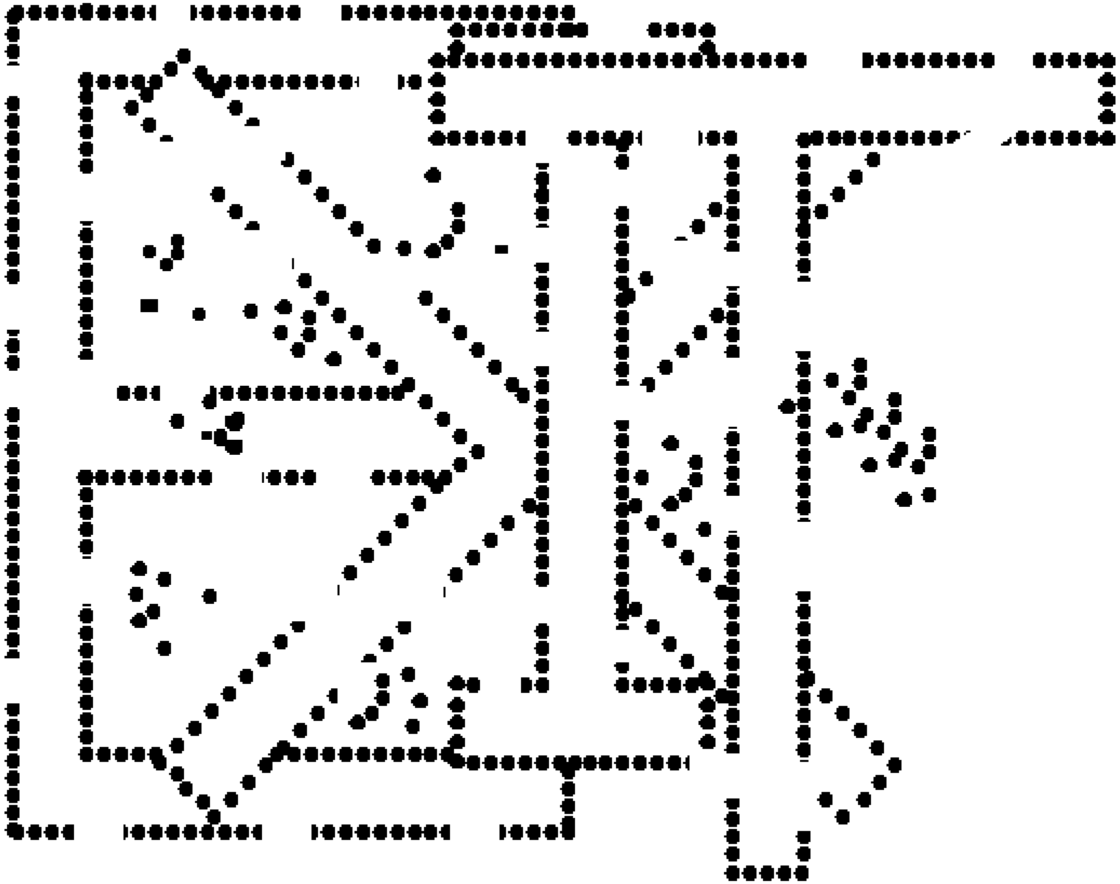
there are problems of integration of sub-functions within a larger function.

E.g. visual perception can operate at different layers of abstraction, where the layers work in parallel influencing one another.

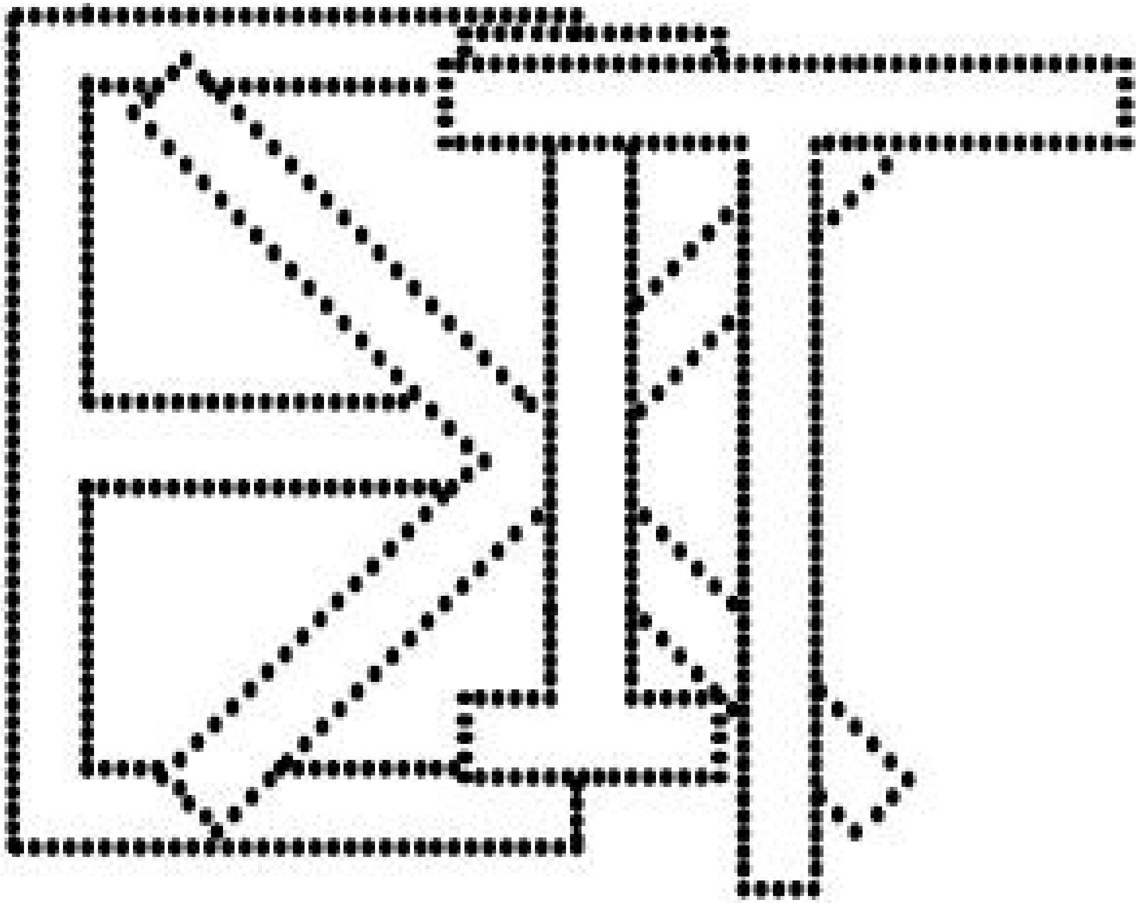
Example:

What word do you see in the next picture?

(Shown for only a short time – e.g. half a second.)



Did you see this?



Multi-layer perception

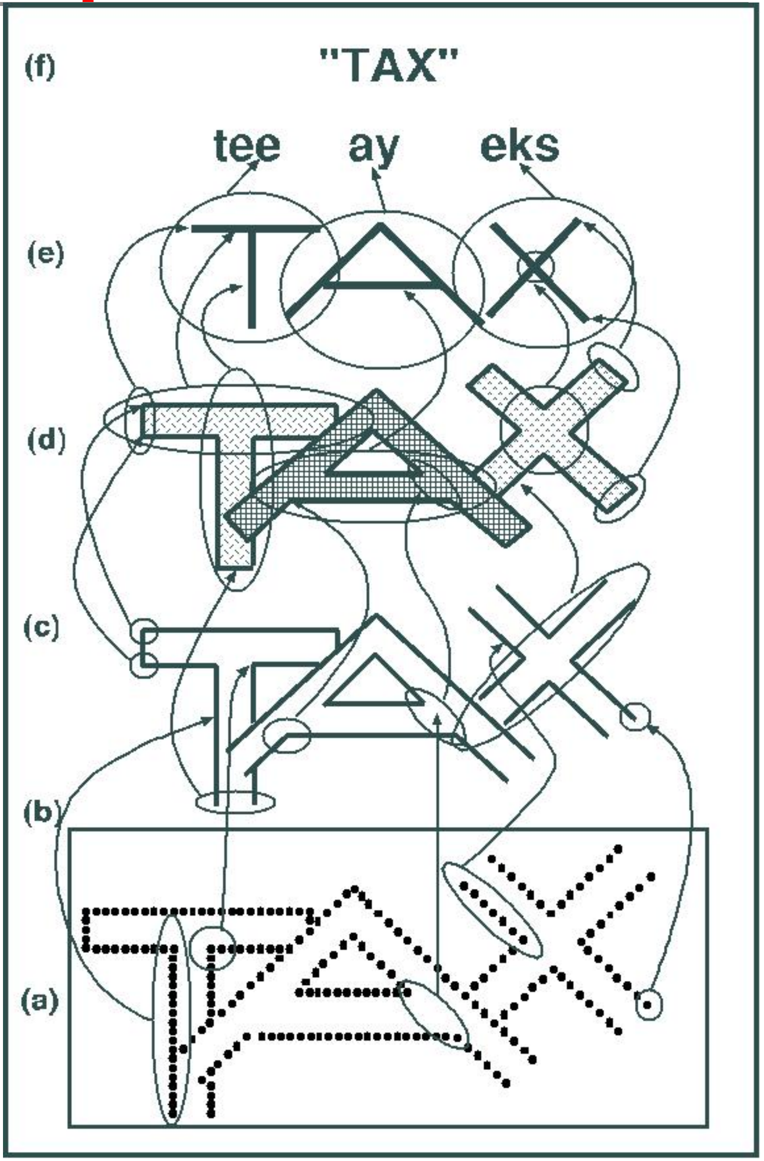
Hypothesis:

Humans looking at the ‘dotty’ pictures of overlapping, outline capital letters simultaneously interpret them at different levels of abstraction, using a mixture of top-down and bottom-up processing.

They use an ontology combining different domains of abstract structures, e.g.

(a) dots and linear dot features
(b) lines and configurations of lines, including junctions of various kinds and pairs of parallel lines
(c) bars (rectangular plates) and laminas made of connected bars, including junctions and other features - in 2.5 dimensions
(d) abstract ‘stroke’ combinations represented by laminas (but also capable of being represented in other ways) including letters (a subset of the stroke configurations)
(e) letter sequences
(f) words (a subset of letter sequences)

The POPEYE program implemented this around 1976. See Ch. 9 of The Computer Revolution in Philosophy 1978



Not only highest-level percepts are outputs

A perceptual system can have many outputs, feeding many different functions in parallel.

Each level of abstraction may be linked to other modules that need the information, e.g. posture control, intruder detection, face recognition, control of manipulation, seeing a problem that needs an explanation, etc.

‘Labyrinthine’ vs ‘modular’ perceptual architectures
Sloman JETAI 1989.

The importance of relations

Relations of different sorts can be perceived, can be thought about, can be used in reasoning or learning.

- **Notions like ‘inside’, ‘overlapping’, ‘between’.**
- **Notions like ‘polygon’ vs ‘triangle’: they differ in parts and relationships.**
- **Some IQ tests and many practical tasks require seeing both relations and relations between relations.**

  E.g. work of Tom Evans on geometric analogy problems, in Minsky’s *Semantic Information Processing* 1968

- **Parts of an object can have different relationships simultaneously: metrical, topological, functional, etc.,**

  How is the difference between a straight wire and a wire with a hook on the end represented?

- **How does the crow represent the goal of getting the hook to lift the bucket?**
- **How is the relation between the hook and the bucket handle represented?**

  Relations between relations: curve of hook related to handle of bucket.

- **What about a sequence of changing relationships – bending a hook.**
- **Joining up two toy trucks?**

See the slides on vision and visual reasoning here
http://www.cs.bham.ac.uk/research/cogaff/talks/

Perceiving structures vs perceiving combinations of features

It is important to understand the difference between

- Categorising
- Perceiving and understanding structure.

  You can see (at least some aspects of) the structure of an unfamiliar object that you do not recognise and cannot categorise: e.g. you probably cannot recognise or categorise this, though you see it clearly enough.

```
 0ooo
 0ooooo-------+
 000ooo000    +
|ooo000ooo----+
+-------------+
```

What is seeing without recognising?

There’s a huge amount of work on visual recognition and labelling e.g. statistical pattern recognition.

But does that tell us anything about perception of structure?
Much work on vision in AI does not get beyond categorisation.

There is something even more subtle and complex than perception of structure.

Perceiving structures vs perceiving affordances

Structures

things that exist, and have relationships, with parts that exist and have relationships

Affordances (positive and negative)

things that could or could not be made to exist, with particular consequences for the perceiver's goals, preferences, likes, dislikes, etc.:

modal, as opposed to categorical, types of perception.

- Betty looks at a piece of wire and (maybe??) sees the possibility of a hook, with a collection of intervening states and processes involving future possible actions by Betty.
- The child looks at two parts of a toy train remembers the possibility of joining them, but fails to see the precise affordances and is mystified and frustrated: presumably he sees the structures because he can grasp and manipulate them in many ways.

How specialised are the innate mechanisms underlying the abilities to learn categories, perceive structures, understand affordances, especially structure-based affordances.

Millions of years of evolution were not wasted!

Affordances and empty space

What are affordances?

An example: How do you see empty space?

Here is a sample:

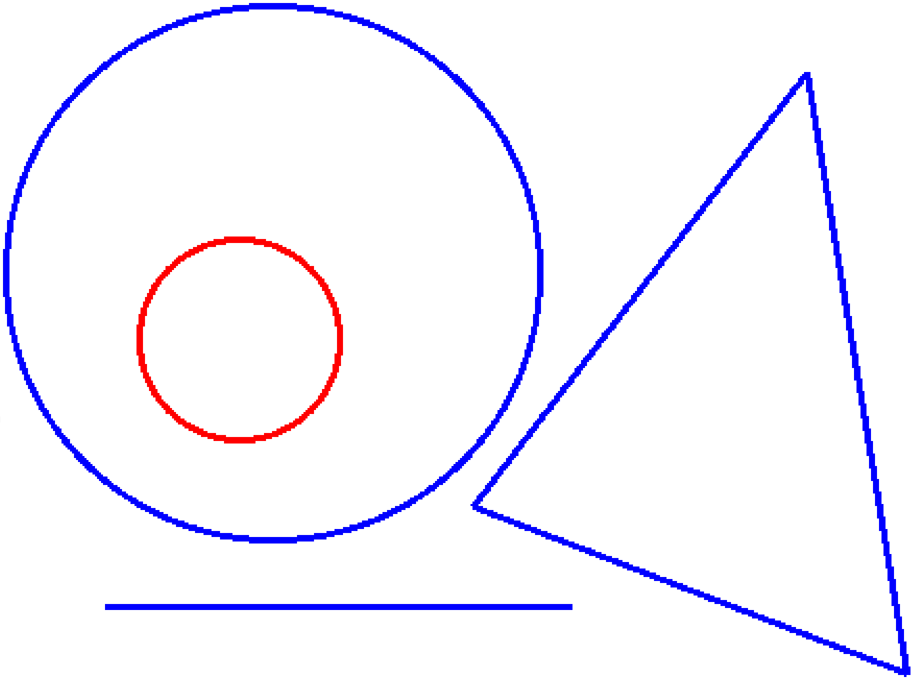
Empty space is a rich collection of possibilities.

(Possible objects, structures, processes, events.)

Affordances and empty space - examples

One way to see space: it affords motion: things that are in one place can move to other places, through (possibly empty) space. This could be important for an agent that moves things.

Consider figures in a Euclidean plane: there are empty spaces to which, and through which, those figures could move, with consequential changes in relationships, including creation of new points of contact or intersection.



- Where will intersections occur if you push the blue line vertically up the page?
- What events will occur if the blue circle moves to the right?
- How many blue circle/triangle intersection points can there be at any time? (Blue circle and blue triangle moving arbitrarily.)
- What changes if the blue circle can grow or shrink?
- If the red circle moves around without changing its size, what is the maximum number of intersection points between circle and triangle? (How do you know you have considered all possibilities?)

Visual reasoning in humans

How do humans answer the question on the right?

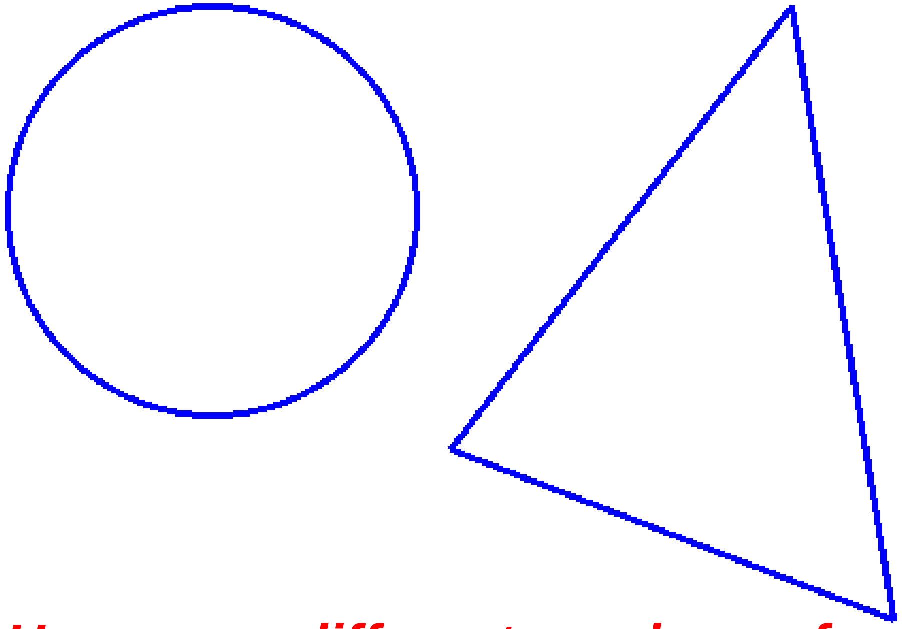
Except for the minority who use logic, this requires the ability to see empty space as containing possible paths of motion, and fixed objects as things that can be moved, rotated and deformed.

Does it require *continuous* change?

Perhaps: but only in a virtual machine!

(To be discussed another time.)

Some people (e.g. Penrose) have argued that computers cannot possibly do human-like visual reasoning e.g. to find the answer ‘seven’ to the question.
(Compare Mateja Jamnik’s PhD thesis)

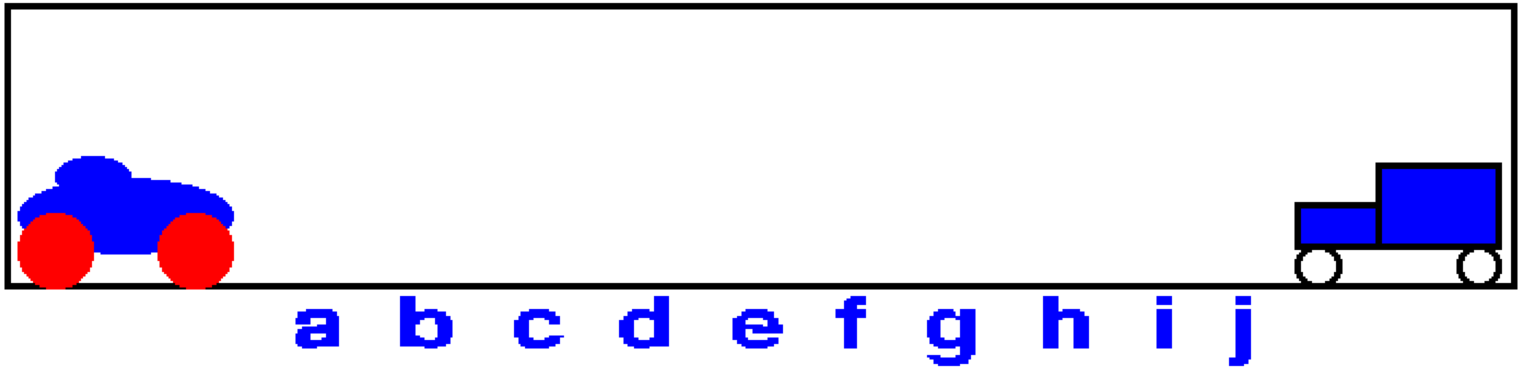


How many different numbers of contact points can there be as circle and triangle slide around, shrink and grow?

Can we find a way to integrate visual perception, imagination, reasoning, problem-solving, language-understanding, action....?

Can we make an artificial mathematician?

Seeing empty space: potentially colliding cars



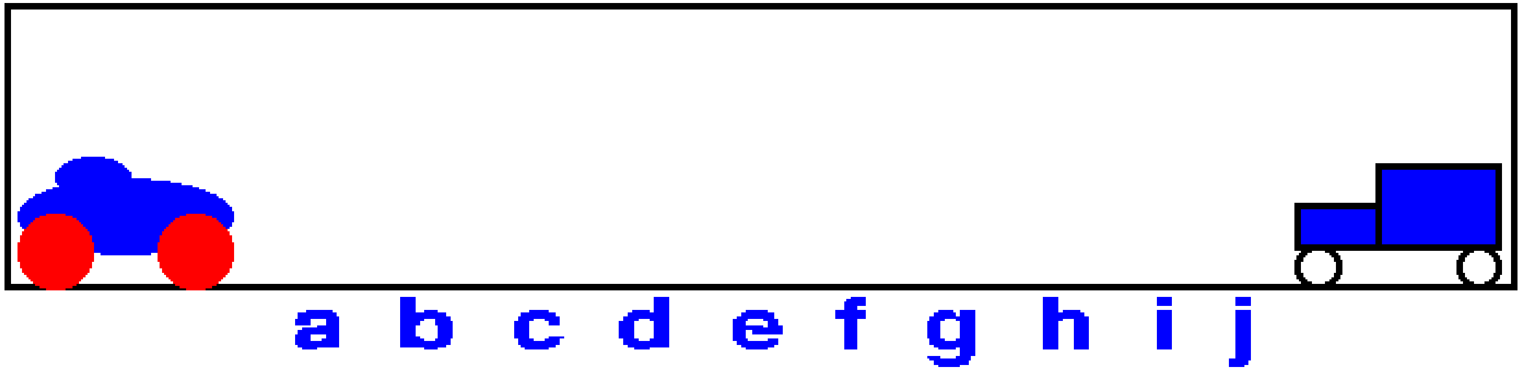
The two vehicles start moving towards one another at the same time.

The racing car on the left moves much faster than the truck on the right.

Whereabouts will they meet?

Where do you think a five year old will say they meet?

Five year old spatial reasoning



The two vehicles start moving towards one another at the same time.

The racing car on the left moves much faster than the truck on the right.

Whereabouts will they meet?

Where do you think a five year old will say they meet?

One five year old answered by pointing to a location near ’b’

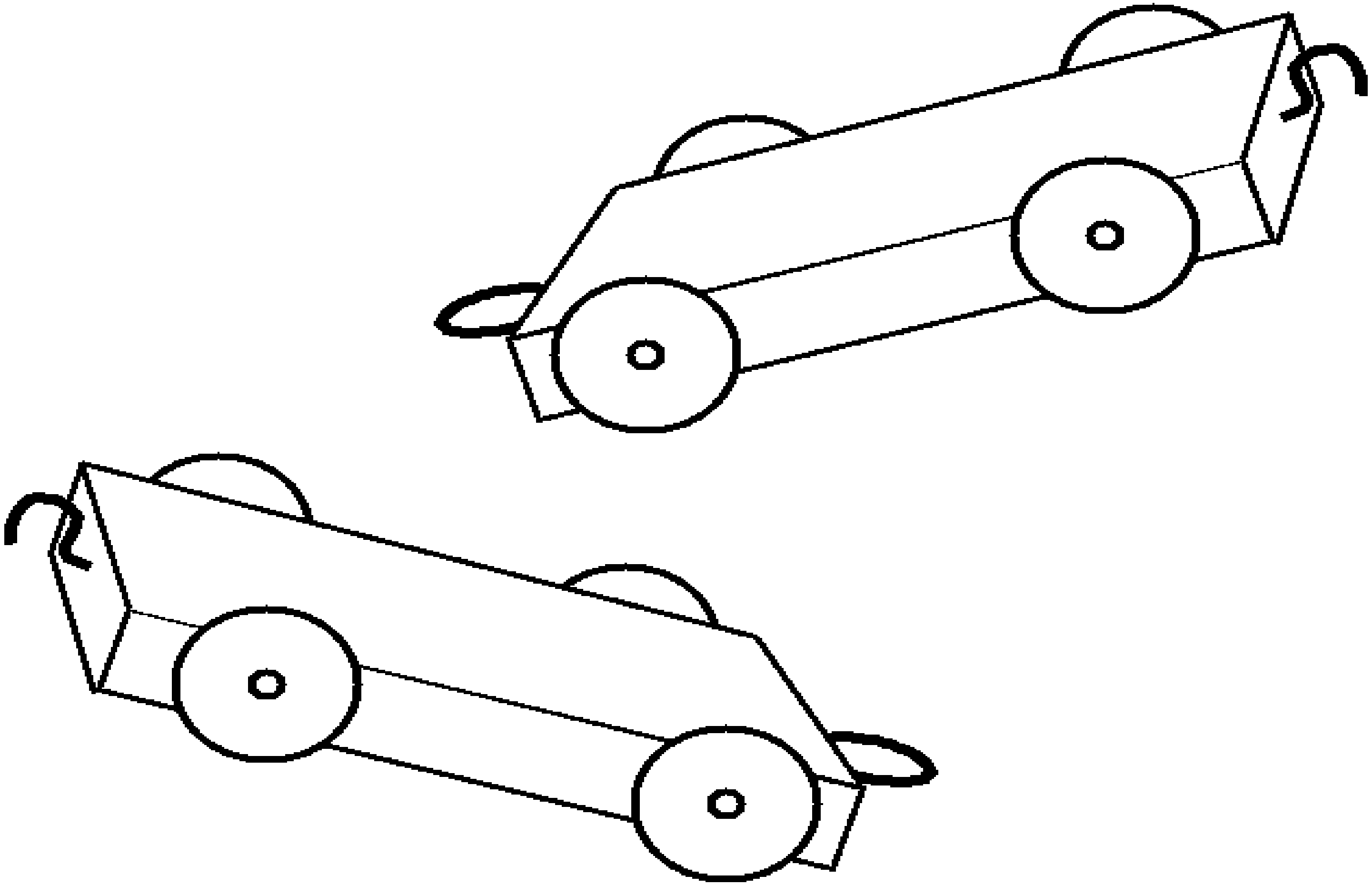
Me: Why?

Child: It’s going faster so it will get there sooner.

What is missing?

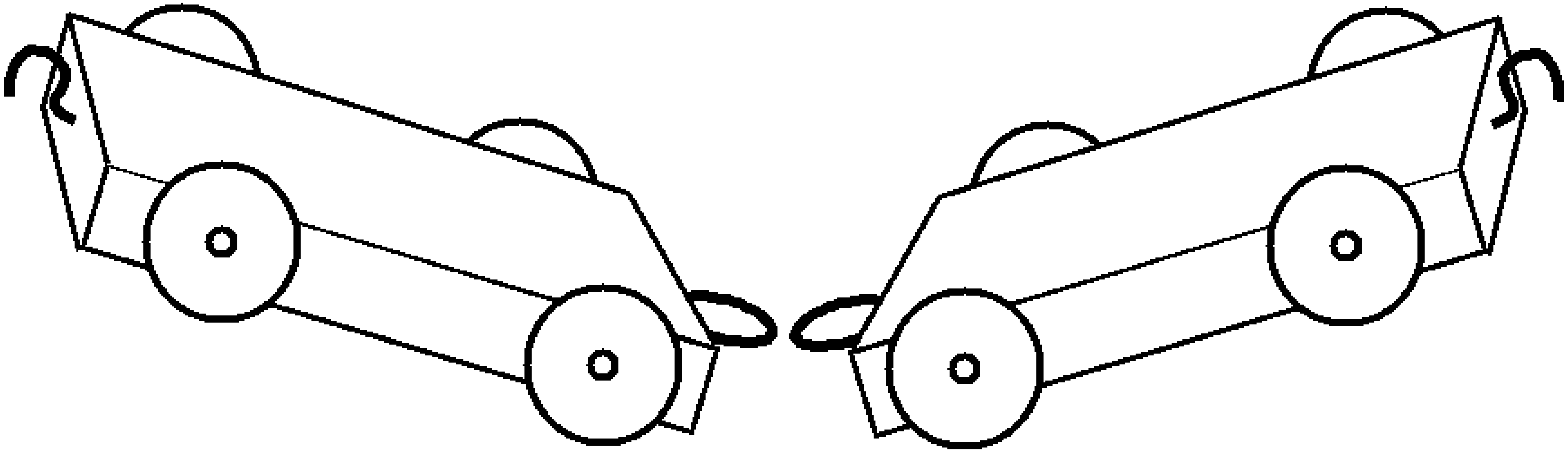
- **Knowledge?**
- **Appropriate representations?**
- **Procedures?**
- **Control? ? ?**

Visualising motion in 3-D space: Building trains



How could you move the trucks to join them together?

Why won’t this work?



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

What changes between a child not understanding why and understanding?

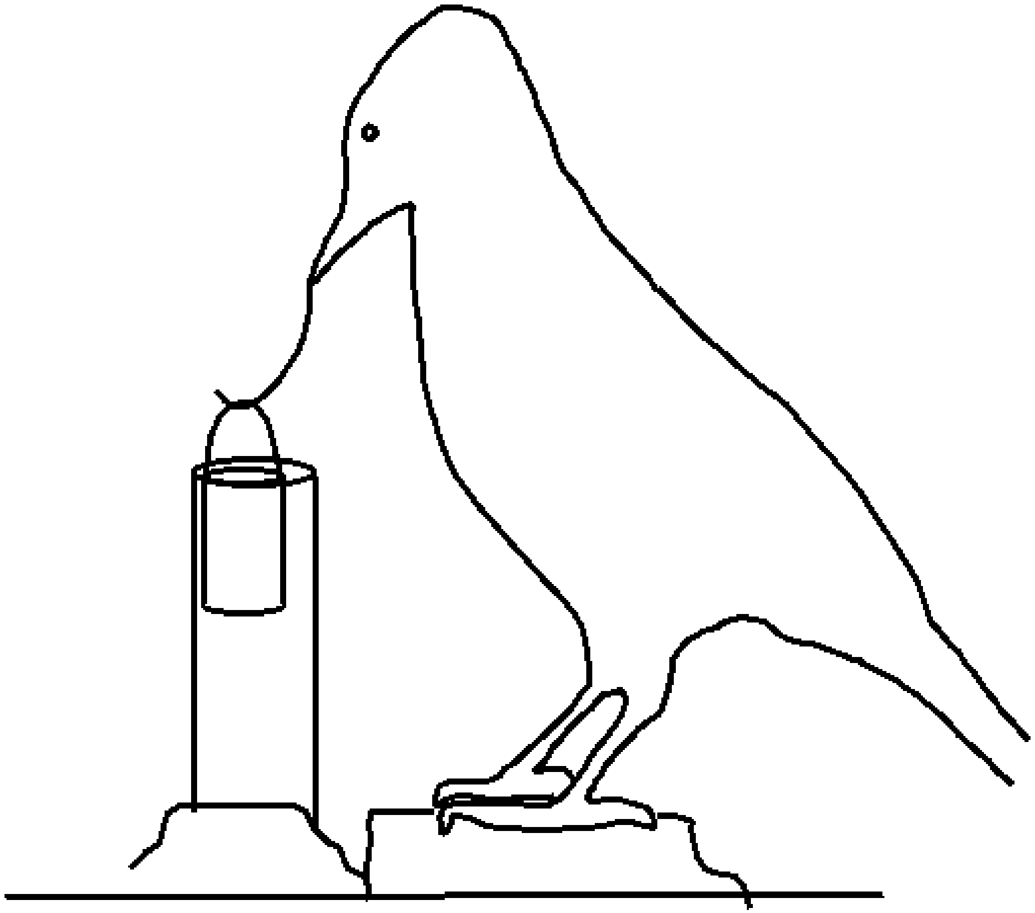
See video of Josh (aged about 18 months) failing to understand the affordances in rings as opposed to hooks:

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

A few weeks later, he seemed to understand.

WHAT MIGHT HAVE CHANGED IN HIM?

Betty



Betty the hook-making crow.

See the video here:

http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm

What are Affordances?

Affordances are not “objective” properties intrinsic to physical configurations.

Rather, they are relational features dependent on the perceiver’s

- **common or likely goals and needs**
- **capabilities for action (physical design + software)**
- **constraints and preferences (avoid stress, injury)**

**However some affordances are relevant to a wide class of agents,
e.g. all those capable of**

- **moving through space,**
- **pushing things,**
- **pulling things,**
- **rotating things, etc.**

This gives some affordances a ‘quasi-objective’ status.

How can affordances be represented?

It is not clear how animal brains represent counterfactual possibilities.

Do they ‘use’ something like modal logics (Steedman 2002)?

Is there some powerful new form of representation waiting to be discovered?

The answer will differ for different sorts of organisms. For purely reactive organisms and machines, the affordances will be implicit in the input-output contingencies that are innate or learnt by the individuals.

For animals and machines capable of thinking ahead, planning, understanding their actions, perhaps affordances in a complex scene can be represented as:

(1) sets of sets of counterfactual conditionals, that are

(2) spatially indexed:

different sets are associated with different parts of objects – a generalised aspect graph.

But this still leaves open what sorts of mechanisms, which forms of representation, and which architectures combining various mechanisms can explain the ability to perceive them, to use them, and to learn new ones.

How do we and how should a robot understand space

Multi-scale multi-purpose spatial understanding

For mathematicians, space is homogeneous: the same in all places and at all scales, but not for most animals, including most humans:

- Manipulable-scale space
- Reachable, mostly visible scale space
- Domestic-scale space
- Urban-scale space
- Geographical-scale space
- Migratory-scale spaces

What is seen is related to possibilities for action. But the possibilities for action are different at the different scales: hence one source of non-homogeneity.

CONJECTURE

Humans and other animals with spatial skills have to learn about all these different aspects of space, location, structure, motion, time, and causation separately (though some aspects may have been ‘learnt’ by evolution and transmitted genetically.)

The integrated view comes much later by a quite distinct sort of learning process, using rare architectural mechanisms.

Asynchronouos interaction

- Many different functions or sub-functions can interact asynchronously.
- E.g. what you see can help as you hear a sentence
- What you see can help you finish your own sentence
- The sentence you hear can help you see
- Seeing something new can help you change your action
- Hearing something can help you change your action
- Seeing someone obeying your instruction can change how you complete the instruction
- Seeing another’s eye gaze can make you change how you finish your sentence

Ordinary life is full of novelty, creativity
(Chomsky, about 40 years ago)

How to think about non-physical levels in reality

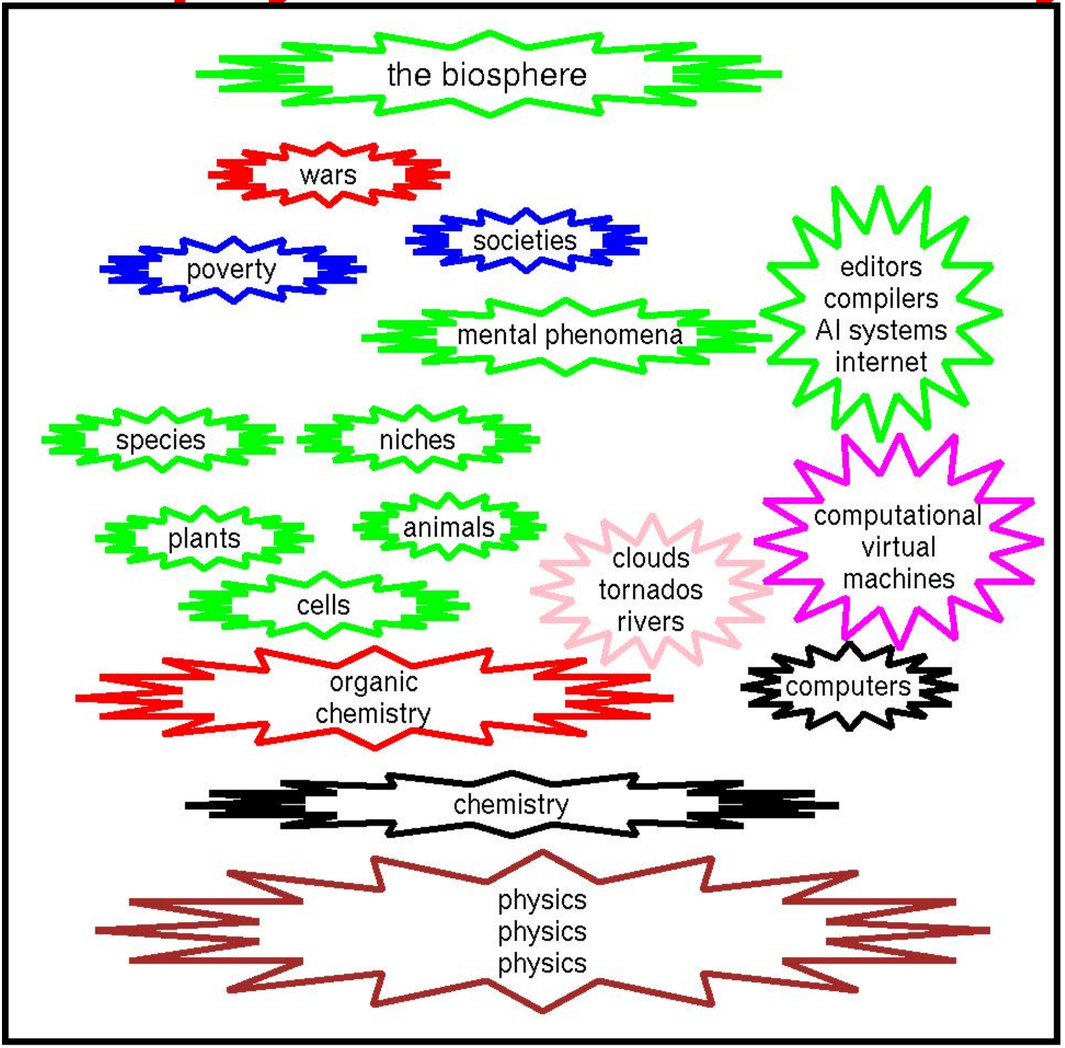
Some philosophers think only physical things can be real.

But there are many non-physical objects, properties, relations, structures, mechanisms, states, events, processes and causal interactions.

E.g. poverty can cause crime.

They are all ultimately implemented in physical systems, as computational virtual machines are, e.g. the Java VM, the linux VM.

Physical sciences also study layers in reality. E.g. chemistry is implemented in physics. Nobody knows how many levels of virtual machines physicists will eventually discover.



See the IJCAI’01 Philosophy of AI tutorial http://www.cs.bham.ac.uk/˜axs/ijcai01/

DIFFERENT VIEWS OF MIND

OLDER APPROACHES:

- A ghost in a machine (dualism)
  - With causal connections both ways: Interactionism
  - With causal connections only one way: Epiphenomenalism
  - With no causal connections: Pre-established harmony
- Mind-brain identity (e.g. the double-aspect theory)
- Behaviourism (mind defined by input-output relations)
- Social/political models of mind
- Mechanical models (e.g. levers, steam engines)
- Electrical models (old telephone exchanges)

PROBLEMS WITH OLDER APPROACHES

- Some lack explanatory power (ghost in the machine)
- Some are circular (Social/Political models of mind)
- Some offer explanations that are too crude to explain fine detail and do not generalise (e.g. mechanical and electrical models)

AI provides tools and concepts for developing new rich and precise theories which don't merely describe some overall structure of mind or mind-body relation, but can show how minds work.

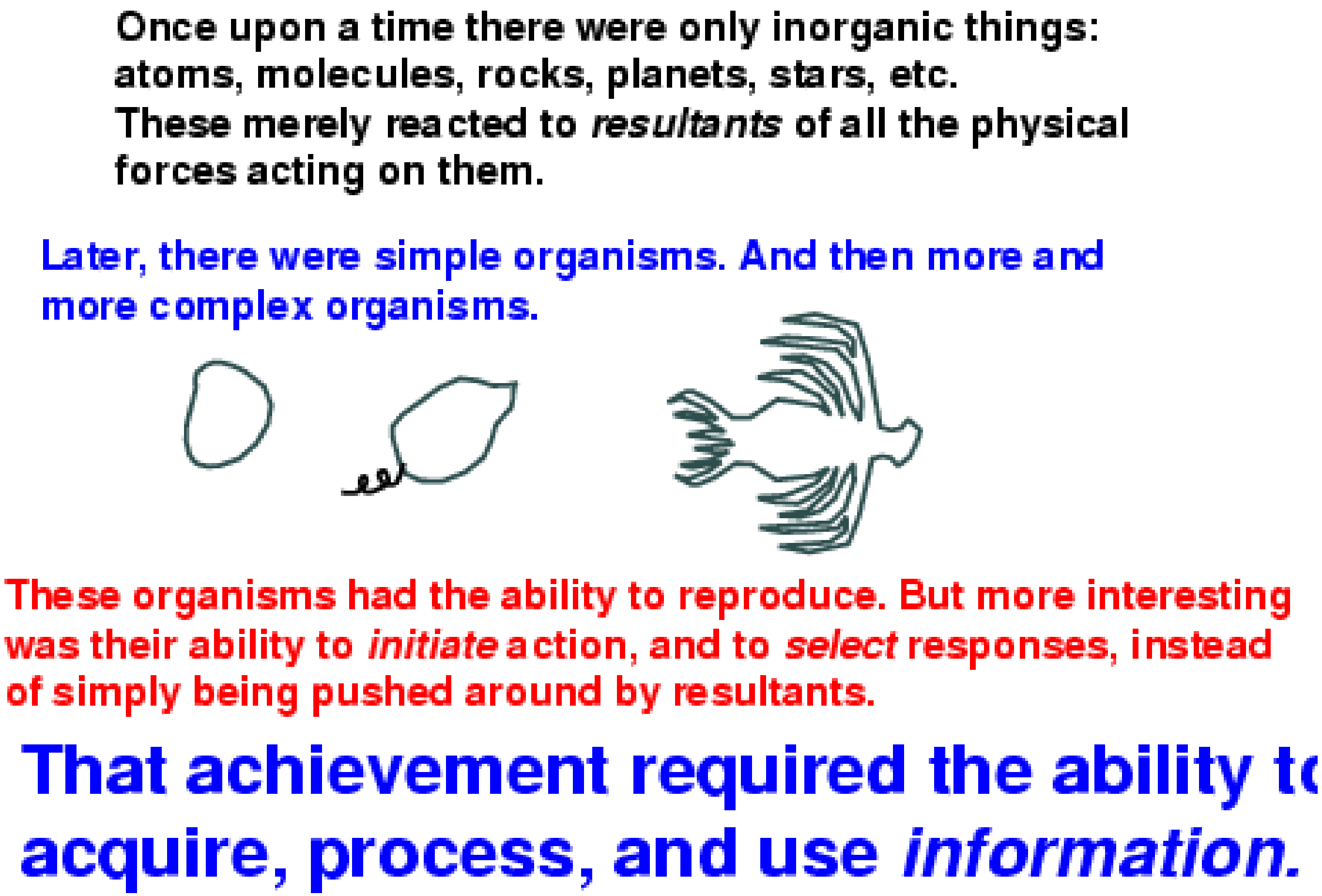
Is there a ghost in the machine?

Every intelligent ghost must contain a machine

If there is a ghost in the machine it requires sophisticated information-processing capabilities to do what minds do.

I.e. there must be a machine in the ghost – an information processing virtual machine – only a virtual machine can have sufficient flexibility and power (as evolution discovered before we did.)

Organisms process information



Once upon a time there were only inorganic things:
atoms, molecules, rocks, planets, stars, etc.
These merely reacted to *resultants* of all the physical forces acting on them.

Later, there were simple organisms. And then more and more complex organisms.

These organisms had the ability to reproduce. But more interesting was their ability to *initiate* action, and to *select* responses, instead of simply being pushed around by resultants.

That achievement required the ability t
acquire, process, and use *information.*

NOTE:

We use “information” in the everyday sense, which involves notions like “referring to something”, “being about something”, “having meaning”, not the Shannon/Weaver technical sense, which is a purely syntactic notion.

Resist the urge to ask for a definition of “information”

Compare “energy” – the concept has grown much since the time of Newton. Did he understand what energy is?

Instead of defining “information” we need to analyse the following:

– the variety of types of information there are,
– the kinds of forms they can take,
– the means of acquiring information,
– the means of manipulating information,
– the means of storing information,
– the means of communicating information,
– the purposes for which information can be used,
– the variety of ways of using information.

As we learn more about such things, our concept of “information” grows deeper and richer.

Like many deep concepts in science, it is *implicitly* defined by its role in our theories and our designs for working systems.

For more on this see http://www.cs.bham.ac.uk/research/cogaff/talks/#inf

Things you can do with information

A partial analysis to illustrate the above:

- You can react immediately (information can trigger immediate action, either external or internal)
- You can do segmenting, clustering labelling of components within a complex information structure (i.e. do parsing.)
- You can interpret one entity as referring to something else.
- You can try to derive new information from old (e.g. what caused this? what else is there? what might happen next? can I benefit from this?)
- You can store store information for future use (and possibly modify it later)
- You can consider alternative possibilities, e.g. in planning.
- If you can interpret it as as containing instructions, you can obey them, e.g. carrying out a plan.
- You can observe the process of doing all the above and derive new information from it (self-monitoring, meta-management).
- You can communicate it to others (or to yourself later)
- You can check it for consistency, either internal or external

... All of this can be done using different forms of representation for different purposes.

What an organism or machine can do with information depends on its architecture

Not just its physical architecture – its information processing architecture.

This may be a virtual machine, like

- a chess virtual machine
- a word processor
- a spreadsheet
- an operating system (linux, solaris, windows)
- a compiler
- most of the internet

What is an architecture?

AI used to be mainly about algorithms and representations.

Increasingly, during the 1990s and onward it has been concerned with the study of architectures.

An architecture includes:

- forms of representation,
- algorithms,
- concurrently processing sub-systems,
- connections between them.

Note: Some of the sub-systems may themselves have complex architectures.

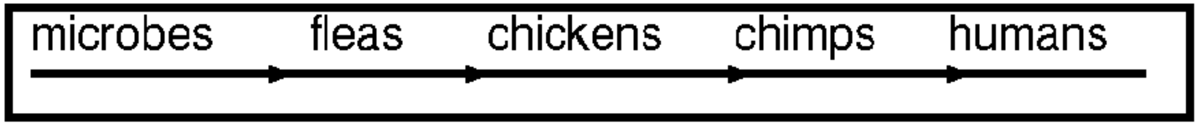
We need to understand the space of information processing architectures and the states and processes they can support, including the varieties of types of mental states and processes.

Which architectures can support human-like emotions?

There’s No Unique Correct Architecture

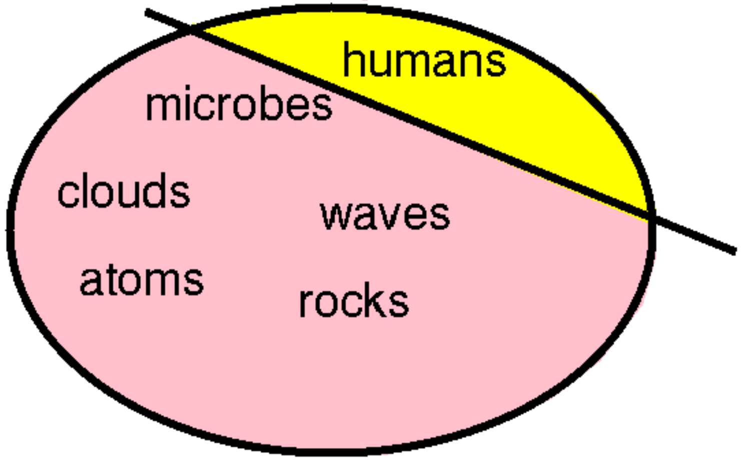
Some tempting wrong ways to think about consciousness:

1. There's no continuum from non-conscious to fully conscious beings



microbes → fleas → chickens → chimps → humans

2. It's not a dichotomy either



humans

microbes

clouds

waves

atoms

rocks

Both ‘smooth variation’ and a single discontinuity are poor models.

We need a better view of the space of possibilities

There are many different types of designs, and many ways in which designs can vary.

Some variations are continuous (getting bigger, faster, heavier, etc.).

Some variations are discontinuous:

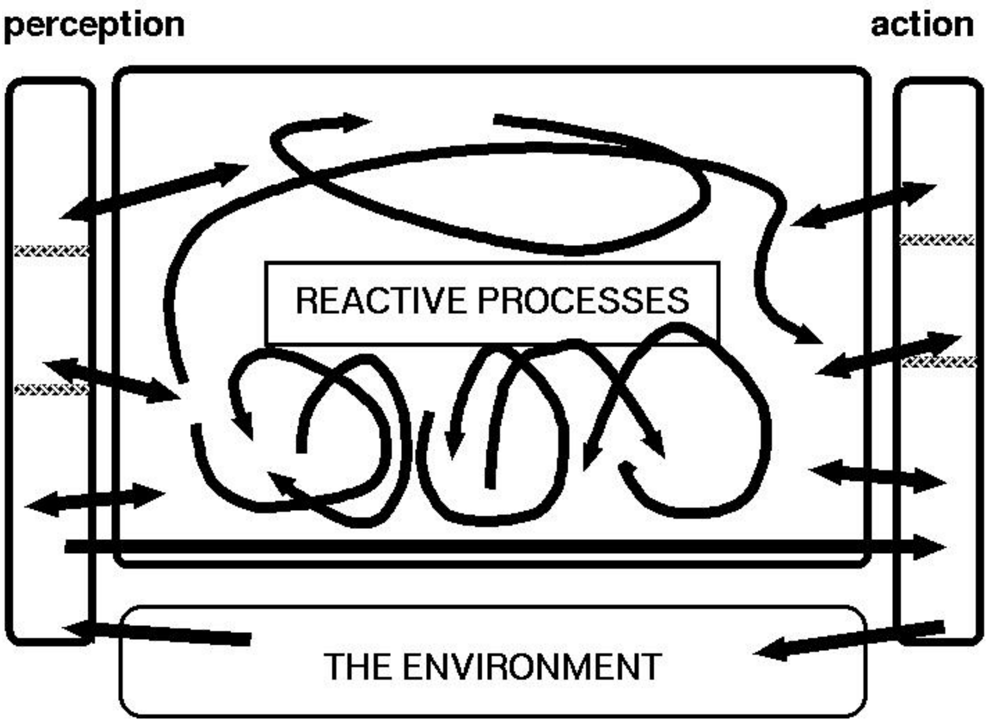
– duplicating a structure,
– adding a new connection between existing structures,
– replacing a component with another,
– extending a plan.
– adding a new control mechanism

Most biological changes are discontinuous — discontinuities can be big or small.

In particular, evolution produces changes of kind as well as degree.

A simple (insect-like) architecture

A reactive system does not construct descriptions of possible futures evaluate them and then choose one. It simply reacts (internally or externally).



An adaptive system with reactive mechanisms can be a very successful biological machine. Some purely reactive species also have a social architecture, e.g. ants, termites, and other insects.

Features of reactive organisms

The main feature of reactive systems is that they lack the ability to represent and reason about non-existent phenomena (e.g. future possible actions), the core ability of deliberative systems, explained below.

Reactive systems need not be “stateless”: some internal reactions can change internal states, and that can influence future reactions.

In particular, reactive systems may be adaptive: e.g. trainable neural nets, which adapt as a result of positive or negative reinforcement.

Some reactions will produce external behaviour. Others will merely produce internal changes.

Internal reactions may form loops.

> **An interesting special case are teleo-reactive systems, described by Nilsson (http://robotics.stanford.edu/)**

In principle a reactive system can produce any external behaviour that more sophisticated systems can produce: but possibly requiring a larger memory for pre-stored reactive behaviours than could fit into the whole universe. Evolution seems to have discovered the advantages of deliberative capabilities.

PROTO-DELIBERATIVE SYSTEMS

- **In a reactive system (e.g. implemented as a neural net) some sensed states with mixtures of features can simultaneously activate two or more incompatibe response-tendencies (e.g. fight and flee).**
- **In that case some sort of competitive mechanism can select between them, e.g. based on the relative strengths of the two seensory patterns, or possibly based on the current context (internal or external e.g. level of hunger or whether an escape route is perceived).**
- **Here two (or more) alternative futures are represented and then a selection is made. Some people call this deliberation.**
- **However, such a system lacks many of the features of a fully deliberative system so we can call it a proto-deliberative system**

FULLY DELIBERATIVE SYSTEMS

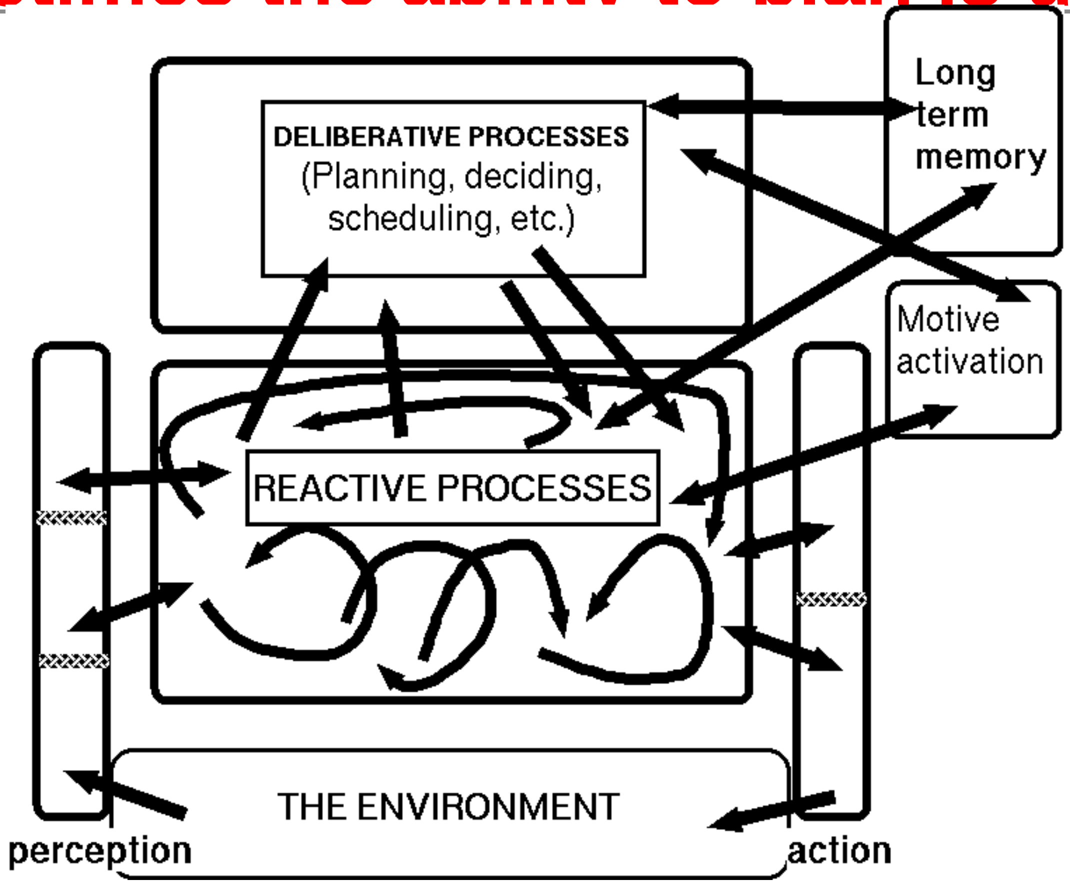
Steps towards fully deliberative capabilities

- The ability to represent what does not yet exist
- The ability to use representations of varying structure
  – Using compositional semantics allows novelty, creativity, etc.
- The ability to use representations of unbounded complexity
  (Contrast fixed size vector representations)
- The ability to build representations of alternative possibilities, compare them, select one

Adding reflective/meta-management capabilities

The ability to be aware of these processes, to categorise, evaluate, modify(reflection/meta-management)

Sometimes the abilitv to plan is useful



Deliberative mechanisms provide the ability to represent possibilities (e.g. possible actions, possible explanations for what is perceived).

Much, but not all, early symbolic AI was concerned with deliberative systems (planners, problem-solvers, parsers, theorem-provers).

Give DELIBERATIVE DEMOS

SHRDLU (pop11 gblocks)

The sheepdog that plans

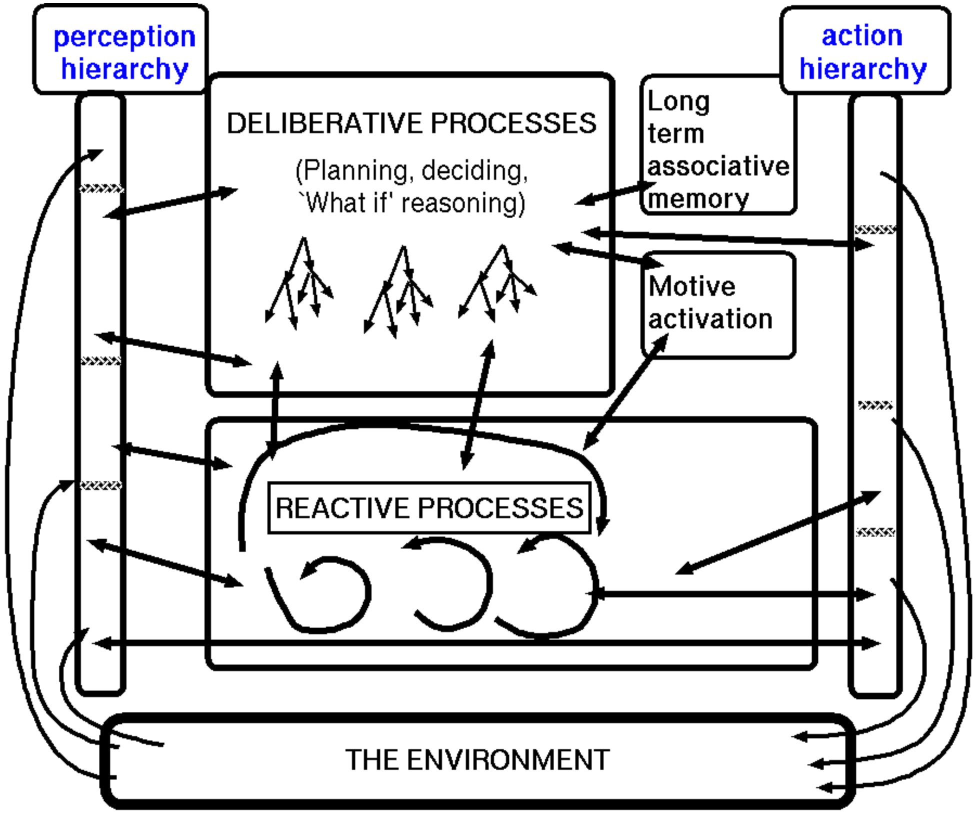
Deliberative mechanisms

These differ in various ways:

- **the forms of representations (often data-structures in virtual machines)**
- **the variety of forms available (e.g. logical, pictorial, activation vectors)**
- **the algorithms/mechanisms available for manipulating representations**
- **the number of possibilities that can be represented simultaneously**
- **the depth of ‘look-ahead’ in planning**
- **the ability to represent future, past, or remote present objects or events**
- **the ability to represent possible actions of other agents**
- **the ability to represent mental states of others (linked to meta-management, below).**
- **the ability to represent abstract entities (numbers, rules, proofs)**
- **the ability to learn, in various ways**

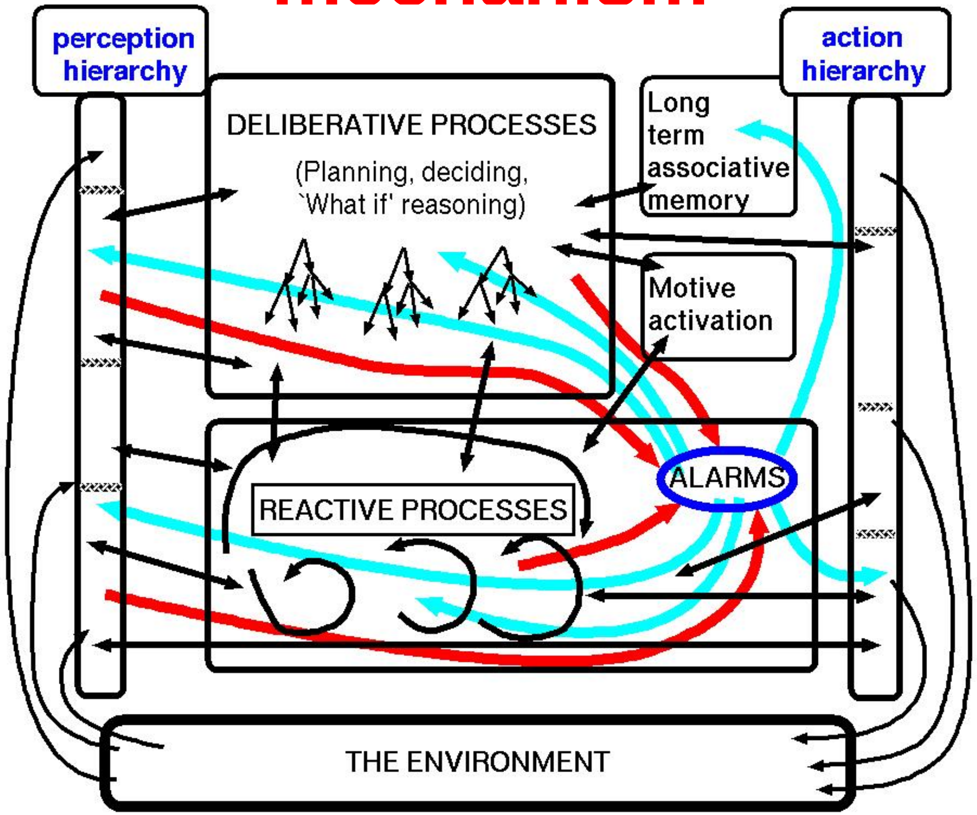
Some deliberative capabilities require the ability to learn new abstract associations, e.g. between situations and possible actions, between actions and possible effects

Evolutionary pressures on perceptual and action mechanisms for deliberative agents



New levels of perceptual abstraction (e.g. perceiving object types, abstract affordances), and support for high-level motor commands (e.g. "walk to tree", "grasp berry") might evolve to meet deliberative needs – hence taller perception and action towers in the diagram.

A deliberative system may need an alarm mechanism



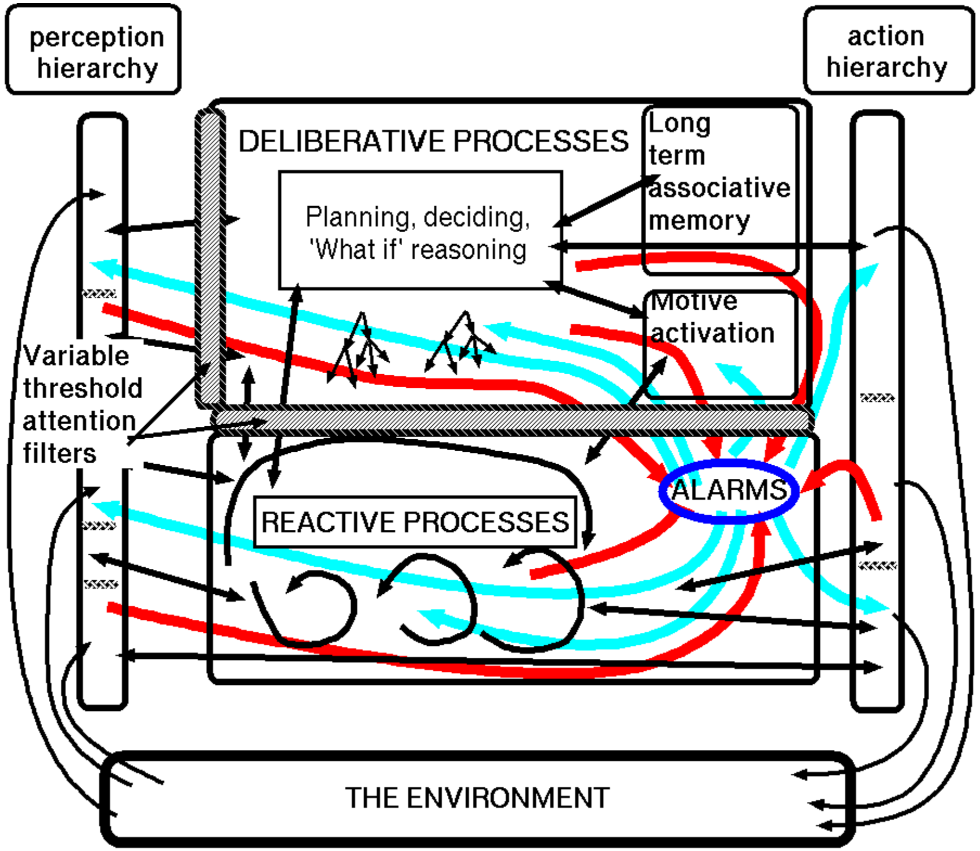
Inputs to an alarm mechanism may come from anywhere in the system, and outputs may go to anywhere in the system.

An alarm system can override, interrupt, abort, or modulate processing in other systems.

It can also make mistakes because it uses fast rather than careful decision making.

A deliberative system may need an alarm mechanism

With some additional mechanisms to act as attention filters, to help suppress some alarms and other disturbances during urgent and important tasks:



Multi-window perception and action

If multiple levels and types of perceptual processing go on in parallel, we can talk about

“multi-window perception”,

as opposed to

“peephole” perception.

Likewise, in an architecture there can be

multi-window action

or merely

peephole action.

Did Good Old Fashioned AI (GOFAI) fail?

It is often claimed that symbolic AI and the work on deliberative systems failed in the 1970s and 1980s and therefore a new approach to AI was needed.

New approaches (some defended by philosophers) included use of neural nets, use of reactive systems, use of subsumption architectures (Rodney Brooks), use of evolutionary methods (genetic algorithms, genetic programming) and use of dynamical systems (using equations borrowed from physics and control engineering).

The critics missed the point that many of the AI systems of the 1970s and 1980s were disappointing partly because they used very small and very slow computers (e.g. 1MByte was a huge amount of memory in 1980), partly because they did not have enough knowledge about the world, and partly because the architecture lacked self-monitoring capabilities: meta-management.

The new emphasis on architectures helps us think more clearly about combining components required to match human capabilities.

The pressure towards self-knowledge, self-evaluation and self-control

A deliberative system can easily get stuck in loops or repeat the same unsuccessful attempt to solve a sub-problem.

One way to prevent this is to have a parallel sub-system monitoring and evaluating the deliberative processes. If it detects something bad happening, then it may be able to interrupt and re-direct the processing.

(Compare Minsky on “B brains” and “C brains” in *Society of Mind*)

We call this meta-management. It seems to be rare in biological organisms and probably evolved very late.

As with deliberative and reactive mechanisms, there are many forms of meta-management.

Conjecture: the representational capabilities that evolved for dealing with self-categorisation can also be used for other-categorisation, and vice-versa. Perceptual mechanisms may have evolved recently to use these those representational capabilities in percepts.

Example: seeing someone else as happy, or angry.

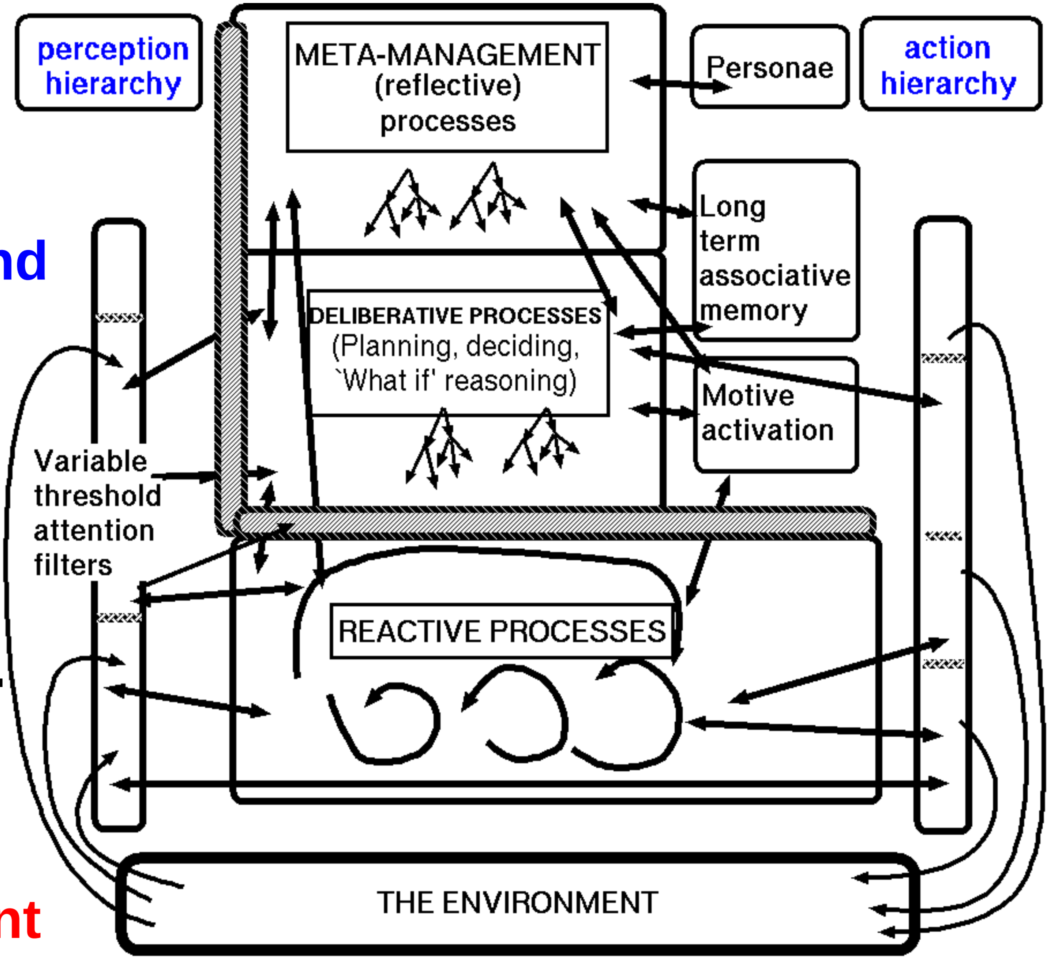
Later, meta-management (reflection) evolved

A conjectured generalisation of homeostasis.

Self monitoring, can include categorisation, evaluation, and (partial) control of internal processes.
Not just measurement.

The richest versions of this evolved very recently, and may be restricted to humans.

Research on 'reflective' AI systems is in progress.


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
REACTIVE PROCESSES
THE ENVIRONMENT

Absence of meta-management can lead to stupid behaviour in AI systems, and in brain-damaged humans.

See A.Damasio (1994) *Descartes' Error* (watch out for the fallacies).

Further steps to a human-like architecture

CONJECTURE:

Central meta-management led to opportunities for evolution of

- **additional layers in ‘multi-window perceptual systems’**

and

- **additional layers in ‘multi-window action systems’,**

Examples: social perception (seeing someone as sad or happy or puzzled), and stylised social action, e.g. courtly bows, social modulation of speech production.

Additional requirements led to further complexity in the architecture, e.g.

- **‘interrupt filters’ for resource-limited attention mechanisms,**
- **more or less global ‘alarm mechanisms’ for dealing with important and urgent problems and opportunities,**
- **socially influenced store of personalities/personae**

All shown in the next slide, with extended layers of perception and action.

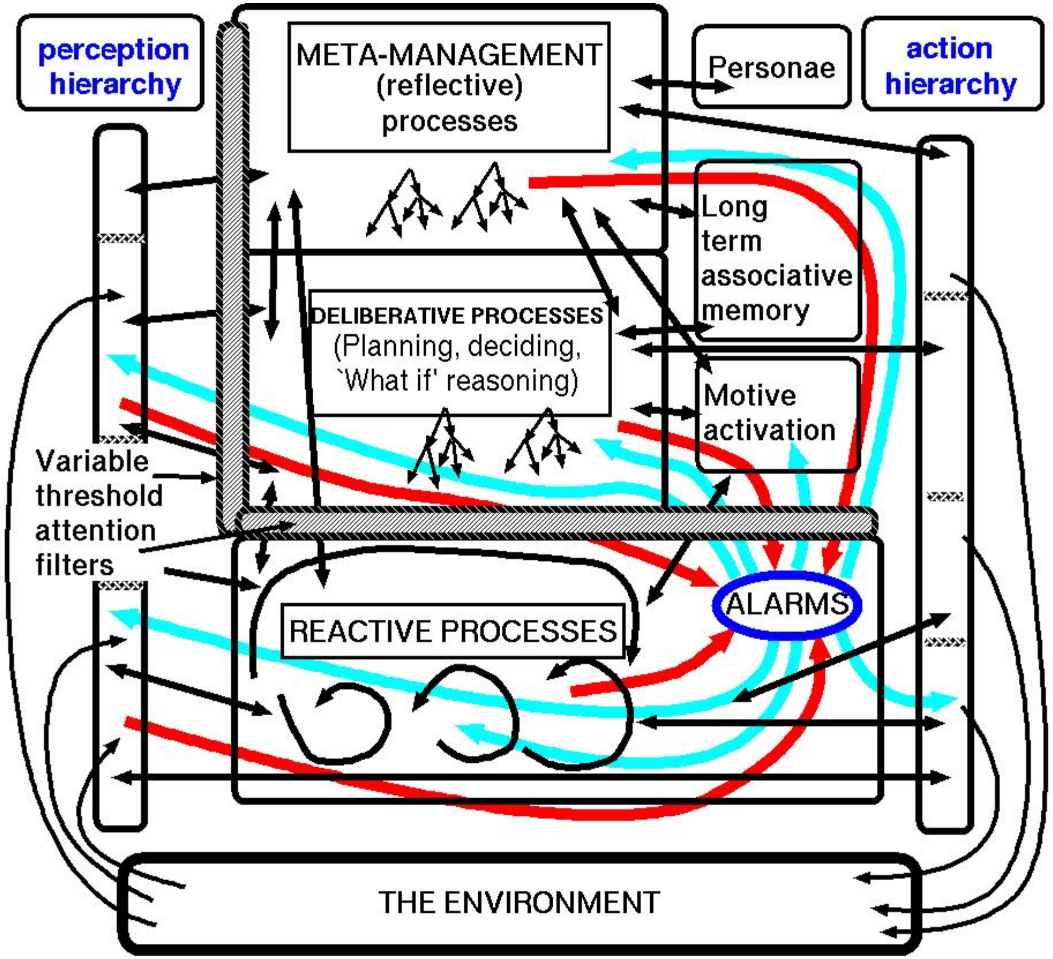
More layers of abstraction in perception and action, and global alarm mechanisms

This conjectured architecture (H-Cogaff) could be included in robots (in the distant future).

Arrows represent information flow (including control signals)

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

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


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

Meta-management systems need to use meta-semantic ontologies: they need the ability to refer to things that refer to things.

Some Implications

Within this framework we can explain (or predict) many phenomena, some of them part of everyday experience and some discovered by scientists:

- **Several varieties of emotions: at least three distinct types related to the three layers: primary (exclusively reactive), secondary (partly deliberative) and tertiary emotions (including disruption of meta-management) – some shared with other animals, some unique to humans. (For more on this see Cogaff Project papers)**
- **Discovery of different visual pathways, since there are many routes for visual information to be used.**
  (See talk 8 in http://www.cs.bham.ac.uk/˜axs/misc/talks/)
- **Many possible types of brain damage and their effects, e.g. frontal-lobe damage interfering with meta-management (Damasio).**
- **Blindsight (damage to some meta-management access routes prevents self-knowledge about intact (reactive?) visual processes.)**

This helps to enrich the analyses of concepts produced by philosophers sitting in their arm chairs: for it is very hard to dream up all these examples of kinds of architectures, states, processes if you merely use your own imagination.

How to explain qualia

Philosophers (and others) contemplating the content of their own experience tend to conclude that there is a very special type of entity to which we have special access only from inside qualia (singular is 'quale'). This generates apparently endless debates.

For more on this see talk 12 on consciousness here http://www.cs.bham.ac.uk/˜axs/misc/talks/

We don't explain qualia by saying what they are.

> **Instead we explain the phenomena that generate philosophical thinking of the sort found in discussions of qualia.**

It is a consequence of having the ability to attend to aspects of internal information processing (internal self-awareness), and then trying to express the results of such attention.

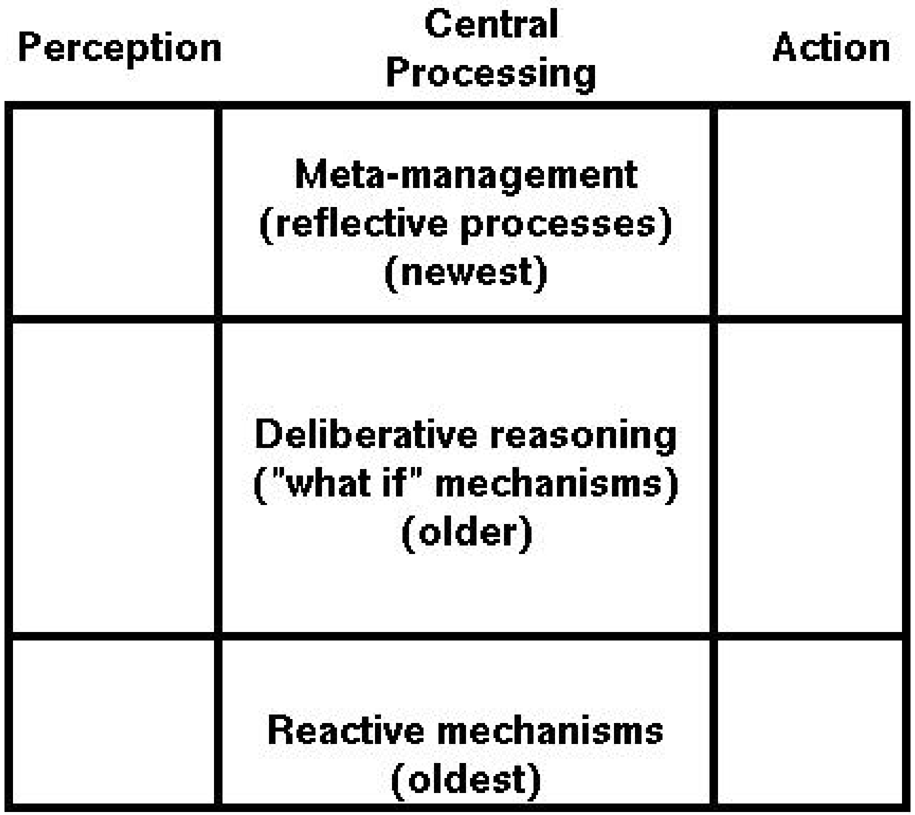
That possibility is inherent in any system that has the sort of architecture we call H-Cogaff, though different versions will be present in different architectures, e.g. depending on the forms of representation and modes of monitoring available to meta-management.

Robots with that architecture may also 'discover' qualia.

How to talk about architectures

Towards a taxonomy (ontology, perhaps a generative grammar), for architectures.

- Types of information used
- Uses of information
- Forms of representation
- Types of Mechanism
- Ways of putting things together in an architecture



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

Architectures vary according to which of the boxes contain mechanisms, what those mechanisms do, which mechanisms are connected to which others.

Also, architectures are not static: some contain contain the ability to grow and develop – new layers, new mechanisms, new forms of representation, new links between mechanisms – e.g. a new-born human's architecture.

Families of architecture-based mental concepts

For each architecture we can specify a family of concepts of types of virtual machine information processing states, processes and capabilities supported by the architecture.

Theories of the architecture of matter refined and extended our concepts of kinds of stuff (periodic table of elements, and varieties of chemical compounds) and of physical and chemical processes.

Likewise, architecture-based mental concepts can extend and refine our semantically indeterminate pre-theoretical concepts, leading to much clearer concepts related to the mechanisms that can produce different sorts of mental states and processes.

Philosophy will never be the same again.

Aristotle: The soul is the form of the body

21st Century: Souls are virtual machines implemented in bodies

We are only just learning how to do backward chaining

- Identifying research goals is hard
- Ordering them so as to select research programmes is hard
- Breaking out of old ways of thinking is hard
- One step is learning to notice things you a have hitherto been ignoring.
- Formulate questions even if you have no idea how to answer them — maybe something will turn up, or someone with different knowledge or abilities will have a great idea.
- We all need to spend more time trying to identify what we don't know.
- Treat everything you read about what others have 'discovered' as provisional, never as established fact: e.g. the researchers may have used the wrong ontology. E.g. claims that
  - there are 'what' vs 'where" visual pathways
  - emotions are needed for intelligence
  - recognition uses collections of features
  - ....

New questions supplant old ones

We can expect to replace old unanswerable questions.

Is a fly conscious? Can a foetus feel pain?

is replaced by new EMPIRICAL questions, e.g.

Which of the 37 varieties of consciousness does a fly have, if any?

Which types of pain can occur in an unborn foetus aged N months and in which sense of 'being aware' can it be aware of them, if any?

Of course, this may also generate new ethical questions, about the rights of robots and our duties towards them.

And that will feed new problems into moral philosophy.

Uncertainty

How to represent this

- Numbers
- Disjunctions
- Partial orders

The representation of uncertainty by probability distributions should be a last resort, as it adds unnecessary complexity in processing.

Wherever possible avoid such things by using more abstract concepts which subsume all the cases: Marr's principle of least commitment

Then constraint propagation can often solve the problem of removing remaining uncertainty.

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

CONCLUSION

Please join the project:
only 300, or maybe 3000, more years to go.

THANK YOU

There’s more on all this in

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