Extended version of slides presented on 12 Sept 2001, based on the paper with the same title in *Proc. British Machine Vision Conference*, 2001, Eds. Tim Cootes & Chris Taylor, Vol 1, pp 313–322.

Evolvable, Biologically Plausible Visual Architectures

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

School of Computer Science
The University of Birmingham

The proceedings paper and related papers can be found at

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

This and other slide presentations can be found at

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

Warning: this is a talk by a philosopher

But one who thinks philosophers should be designers

(as you'll see).

This is a sequel to:

A. Sloman, 'On designing a visual system (Towards a Gibsonian computational model of vision)', in *Journal of Experimental and Theoretical AI*, vol 1, no 4, pp. 289–337, 1989

http://www.cs.bham.ac.uk/research/cogaff/81-95.html#7

That in turn, is a sequel to the sections on vision in my out of print 1978 book,

The Computer Revolution in Philosophy: Philosophy Science and Models of Mind.
This is now online:
http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html

See also

Shimon Ullman, *High-level vision: Object recognition and visual cognition*, MIT Press, 1996.

That book makes some similar points.

I have recently proposed a (partly) new theory of vision as process simulation, described in this PDF presentation: http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0505

The functions of vision

If we wish to understand real visual systems we must try to understand what animals, including humans do with vision.

This may go far beyond what your first thoughts about the functions of vision are.

E.g. you may think that vision is used

- to compute a depth map
- to tell you about distances, orientations, shapes, colours, textures of visible surfaces in the scene.
- to segment and classify objects in the environment
- to control what you should do next

It is all that and much more. And most of what human vision does goes far beyond what current AI/Robotic systems can model and far beyond what current theories of brain mechanisms are able to explain.

So this talk is about some of those functions, and about some ideas relevant to producing adequate models and theories in the future.

Vision is about awareness of what's going on

Show video of child with trainset and tunnel:

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

- Notice how the child aged about 32 months is aware of what's going on around him even when he can't see everything e.g. because something is behind him or because it is out of sight in the tunnel.
- He does not think the train gets smaller as it is pushed into the tunnel and is not surprised when the invisible bit emerges from the far end.
- He clearly sees things that continue to exist while unperceived, and when the back of his head knocks over a toy tree he knows what has happend and how to move to get into a position to fix it.
- All the time his visual system is rapidly sampling different bits of the environment (active vision), but there is no reason to believe that with each switch of gaze he starts all over again building a model of what's going on around him.
- Vision is not a source of information about what is in the current retinal image: rather, it is a source of information about what is in the environment.
- The constantly changing retinal image, along with constantly changing tactile and auditory information all contribute to that ongoing percept.

  Compare J.J. Gibson, (1966). *The Senses Considered as Perceptual Systems.*

KEY IDEAS

Vision is about

- processes in the environment
- structures in the environment
- relationships in the environment
- causes and effects in the environment
- opportunities in the environment
- obstacles and constraints in the environment
- what is likely to happen in the environment
- what the perceiver is doing in the environment
  (including failing to do, nearly succeeding, etc.)

Much of this is about what J.J.Gibson called 'affordances' (positive and negative) for the animal or robot.

See his 1979 book. *The Ecological Approach to Visual Perception*

Mechanisms able to acquire and process such information may need

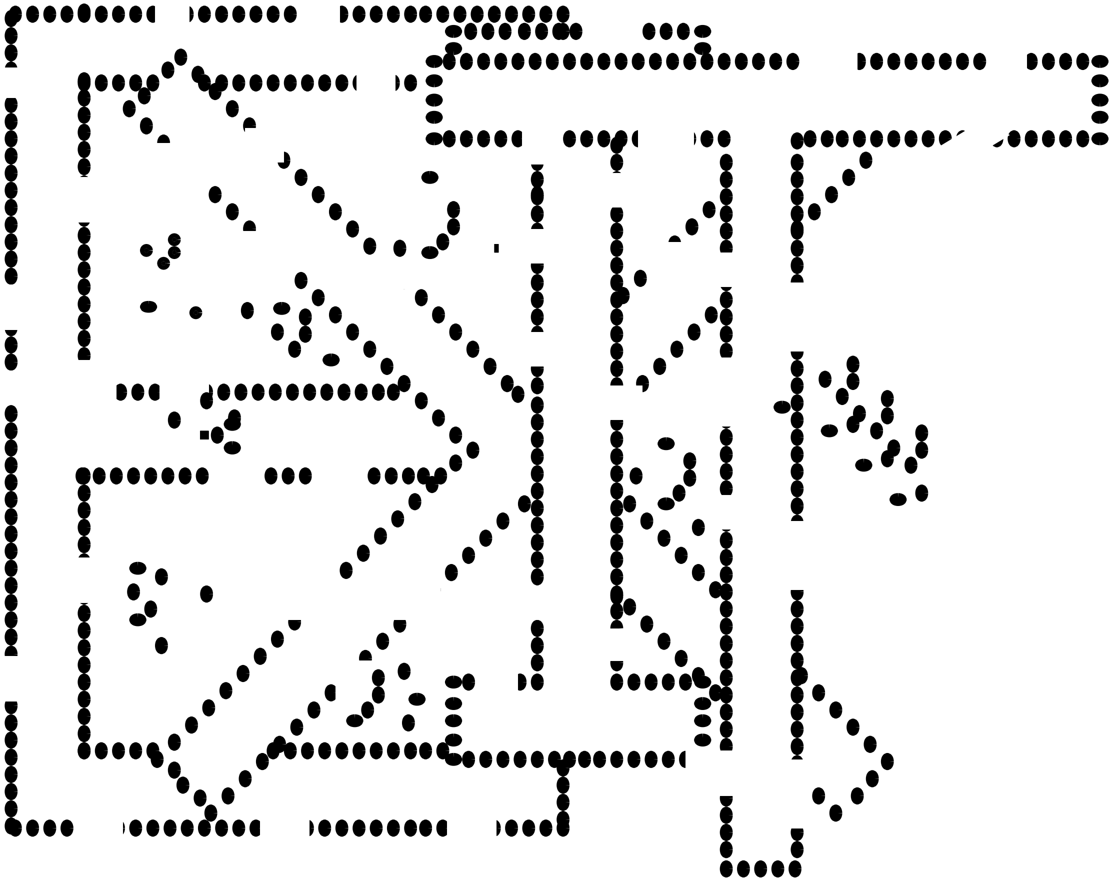
- Different levels of processing to occur in parallel
- Different forms of representation
- Different ontologies
- Different background knowledge about the environment

An example of YOUR visual system at work

How quickly can you recognize the next word?

Allow yourself about half a second for the next slide then move on.

What word do you see?



What did you see?

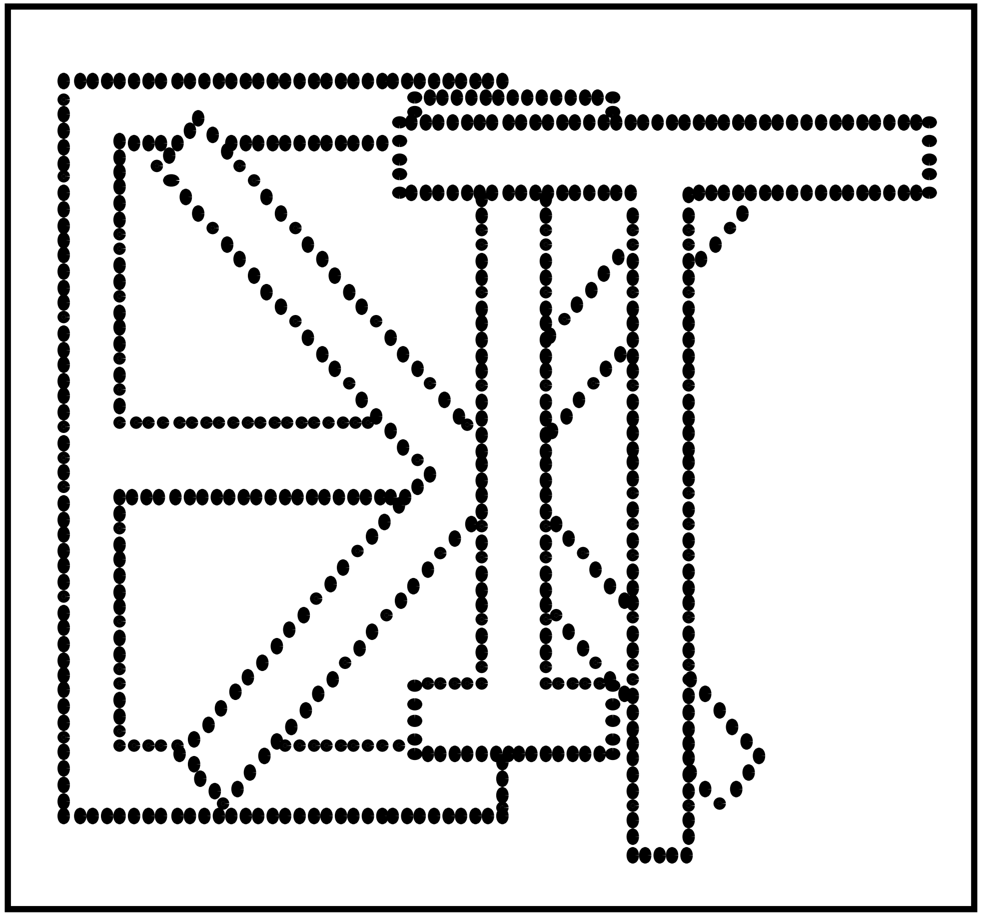
If you did not see a word, try going back for a slightly longer period.

If you did see a word carry on to the next slide.

Some work done in the 1970s: POPEYE

The Popeye project (using POP2, a precursor of Pop-11) investigated how it is possible for humans to see structure in very cluttered scenes, where structure exists at different levels of abstraction.

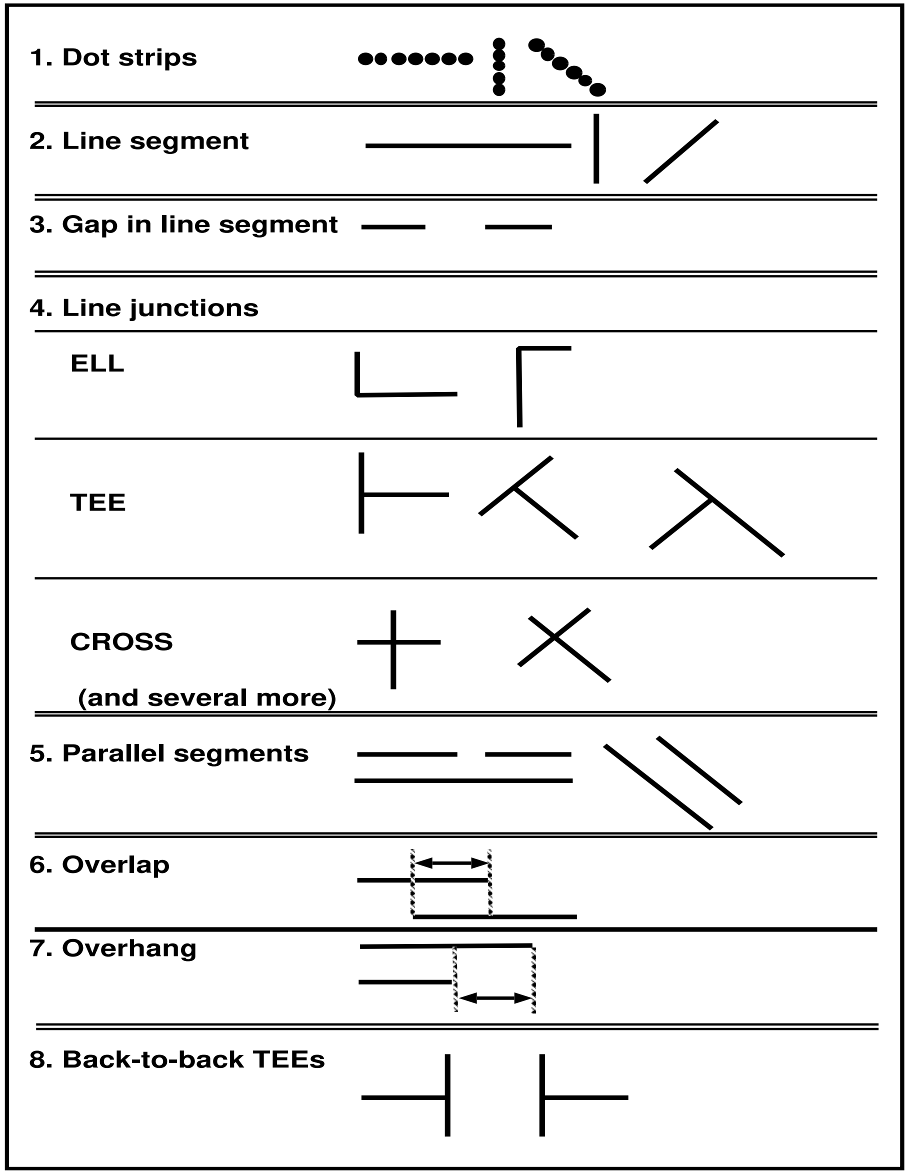
The program developed was able to process a noisy image at different levels of abstraction, using a mixture of concurrent bottom-up and top-down processing, as a result of which it worked quickly and reliably in easy cases, and, a bit like humans, degraded gracelly as noise and clutter increased.



See *The Computer Revolution In Philosophy* (1978) Chapter 9
http://www.cs.bham.ac.uk/research/cogaff/crp

An ontology for seeing Popeye's dotty pictures

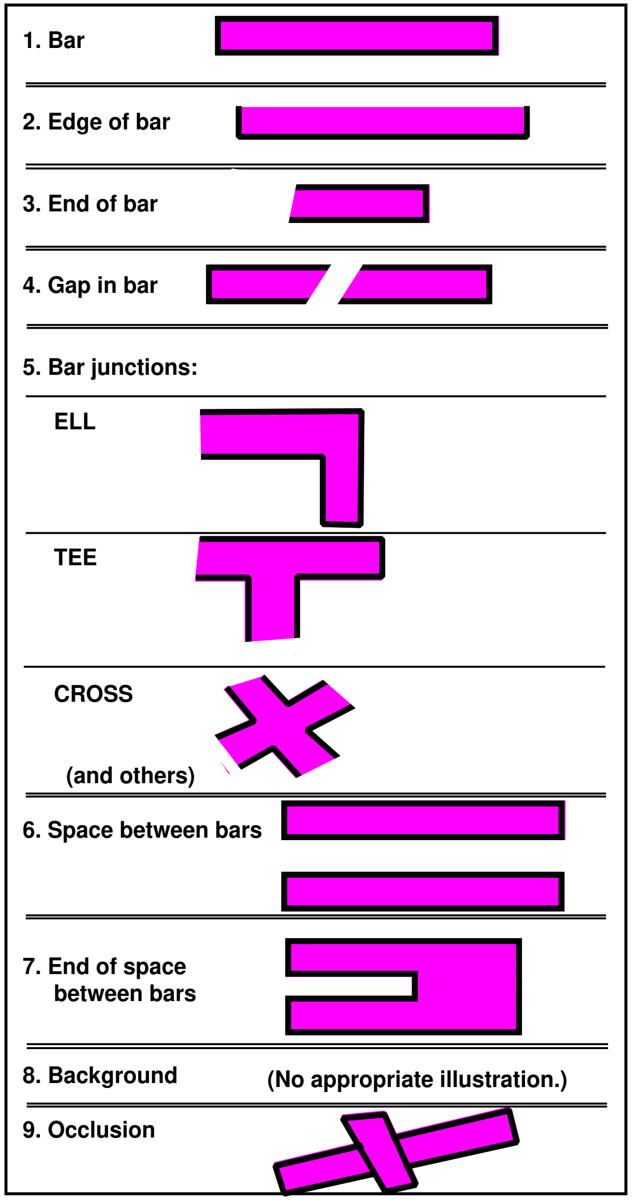
Useful fragments at different levels of abstraction



Parts of the lamina ontology

Some of the significant fragments detectable in the domain of overlapping laminas.

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



Larger scale line ‘phrases’

It is useful to know about frequently occurring fragments, as we do with linguistic fragments (e.g. ‘up the hill’, ‘for the sake of’, ‘in the way’, ‘under the table’).

See J.D. Becker on ‘The phrasal lexicon’

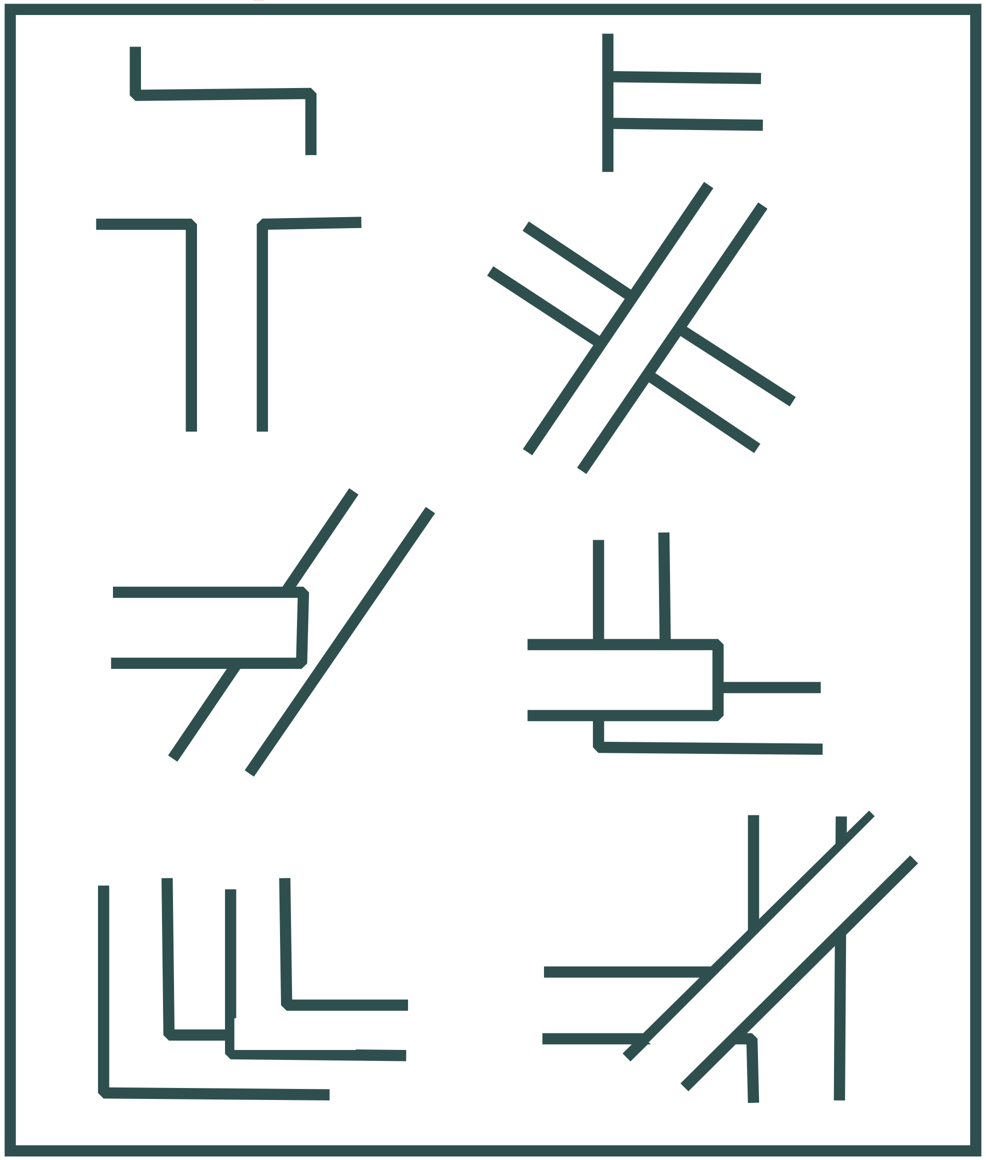
Likewise knowing about familiar objects and fragments of objects in the environment may help visual processing

So recognition is not just about complete objects.

Larger “phrases” in the “language” of line fragments.

Could a neural net learn such things?

Are there any known mechanisms that are appropriate?



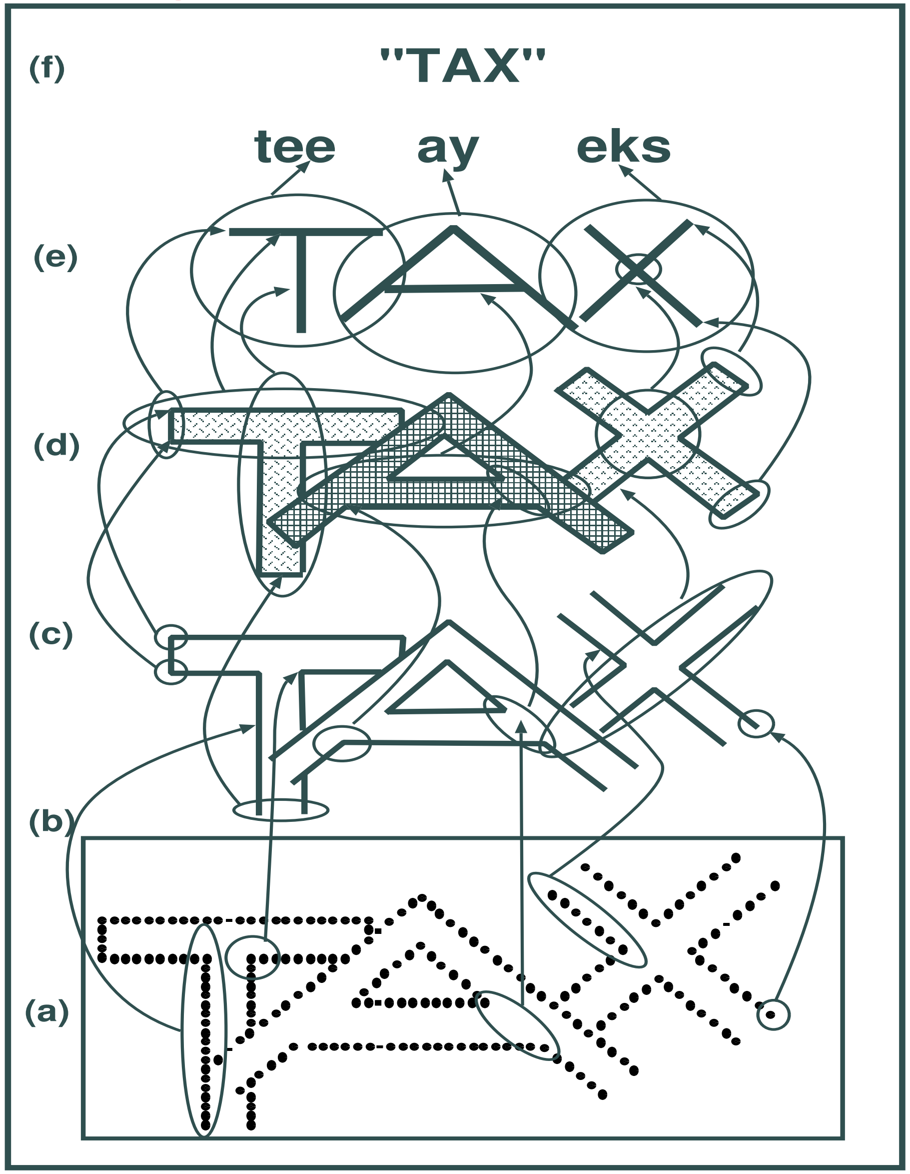
Putting it all together

The Popeye architecture specified concurrent processing at all these different levels of abstraction.

Sub-systems at different levels could interact with higher- or lower-level 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!



For a discussion of the need to extend perception of multi-level stuctures to perception of multi-level processes see http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0505

Building a self-contained visual system is not enough

The POPEYE system could identify components in the scene and their relationships and use that to guide the recognition of other components and relationships, at different levels of abstraction.

(Critics tended to confuse this with the then fashionable notion of "heterarchic" processing strongly criticised by David Marr.
See http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html)

But vision does not occur in isolation.

A visual system is part of a whole organism, or robot.

What the visual system needs to do will in part depend on what the organism needs, and on what other components there are in the system.

Other components can ask the visual system questions, can use information provided by vision, can help to train the visual system, can provide information for the visual system, ...

Vision is related to goals and actions

So far we have introduced the idea that vision (and perception in general):

- deals not only with recognition of individual objects, but also with structures (parts with relationships, where the parts have parts with relationships, etc.)
- where the structures can involve different ontological levels (different levels of abstraction: dots, lines, laminas, strokes, etc.)
- where the perceptual processes can involve bi-directional influences, not just bottom-up data-driven processes

But vision also involves changes:

- moving objects
- actions produced by the perceiver, who needs to control those actions
- possible future events
- possible past events (which might explain something seen now)

J.J.Gibson:

Organisms need to see not only what is there, but also the positive and negative affordances, i.e.

what could or could not happen or be done that might be relevant to the perceiver's goals, needs, preferences, interests, ...

So

Vision, and other forms of perception should not be studied in isolation from other kinds of functions of a complete robot or organism, and should be related to the study of different kinds of knowledge, different sorts of concepts, different kinds of representations, and also different aspects of the environment, some of which are physical and some of which are far more abstract, including affordances, which vary from one kind of perceiver to another.

AI USED TO BE MORE INTEGRATED

In earlier days, e.g. 1960s, 1970s and early 1980s, people working on sub-problems

(e.g. language, planning, reasoning, learning, vision, motor control, etc.),

whether they intended to do science or to do engineering,

knew that what they were doing was part of a larger task:

understanding principles relevant to designing and implementing complete[*] working systems containing all the components working together.

They learnt about other sub-fields because they all went to the same conferences, e.g. Machine Intelligence, IJCAI, AISB, AAAI, ECAI

There weren't many other AI conferences, in those days!

[*] Note: not all complete systems are equally rich – there are "toy" complete systems!

As AI grew more popular, it fragmented

More and more people got involved in AI.

So, inevitably, the field grew more and more fragmented,

into sub-fields where people work on narrowly focused problems and techniques.

Even sub-fields have become fragmented:

sub-sub-fields are full of hard problems and more and more complex and specialised techniques are being developed for dealing with them.

As a result there is very little interest in how to put things together.

Everyone (almost) is too busy with more focused problems.

And most researchers don’t know much (or care much?) about what researchers in other fields are doing.

That’s fine ...

If your objective is only to solve precisely specified and suitably narrow practical problems — a worthy engineering goal.

Moreover, much of the detailed work can also contribute to the design of mechanisms required in fully functioning integrated architectures.

My aim, however, is conservation of a rare species — preventing extinction of the subset of people interested in putting it all together!

Can we re-assemble AI?

VISION IS CRUCIAL: THE HARDEST PROBLEM IN AI
(and psychology, neuroscience, ...)

Will we ever be able to design machines with visual and other capabilities of squirrels, gibbons, or magpies (let alone humans)?

First, we need to understand what those visual capabilities are: which may be far from obvious.

Identifying the full range of human visual capabilities is harder than it seems, since we don’t always know when we are using visual capabilities – as explained below.

The Good News

Over the last decade another sub-activity has grown up:
the study of architectures.

Previously, the three main kinds of AI research, going back 50 years, were:

- The study of forms of representation
- The study of algorithms for performing various kinds of computations over those representations.
- The study of factual and procedural domain-specific knowledge to be encoded in representations and algorithms.

  (e.g. knowledge of stereo, of lighting and the optical properties of surfaces, of the image formation process, and much procedural know-how)

The study of architectures investigates ways of putting these things together.

Chaos in architecture-land

Unfortunately, there is much confusion in discussions of architectures.

E.g. different people use apparently similar diagrams and descriptions, to refer to different architectures.

E.g. “multi-layer” architecture means different things to different people. (Compare our three layers below.)

There is also too much factionalism (narrow vision).

Many people commit themselves to one or other type of mechanism (e.g. neural nets) or one type of architecture (e.g. subsumption) without having any really clear idea what the alternatives are or what the trade-offs between them are, ignoring the history of the field.

Some also teach their students to be too narrow-minded — they grow up knowing only one way to think!

Contrast Minsky’s analysis of trade-offs between neural and other forms of computation – what’s best depends on what the problem is:

‘Future of AI Technology’, 1997,

http://www.media.mit.edu/people/minsky/papers/CausalDiversity.html

Original version in Toshiba Review, Vol.47, No.7, July 1992.

Another problem: what needs to be explained?

It is too easy to assume we know what capabilities need to be explained, for they are our capabilities.

Problems with this assumption:

- We are not necessarily aware of *which* capabilities we use in many tasks, or even *that* we are performing them, e.g. posture control, recognising features, analysing structures, solving image correspondence problems, reacting to facial expressions, doing visual learning.
- In particular, we may not always be aware of the role of visual processing in some of those tasks, e.g. in doing abstract mathematics
  (See Talk 7 here: http://www.cs.bham.ac.uk/~axs/misc/talks)
- What may appear to be *one* task, e.g. estimating distance, or seeing shape, or comparing angles, may actually be different tasks in different contexts, performed in different ways in different parts of the information processing architecture, using different forms of representation, e.g. judging distance in preparing to jump across a ditch, and judging distance in selecting a plank to lay across the ditch.

We still need to identify the diverse functions of vision: a requirement for building adequate explanatory theories or working models.

In humans there is great diversity of visual capabilities

E.g.

- What we see goes far beyond geometric/spatial structures and properties.
- We see many things that are abstract, some of them possibly shared with other species, others unique to humans.
- We can train ourselves to see and interpret things more quickly and fluently (e.g. learning to sight-read music, learning to play tennis).

Some human visual capabilities are culture-specific or location specific (e.g. in snow or in forests), while others are more general, e.g. the ability to see symmetries.

Non-geometrical visual percepts are harder to explain:

– seeing causal and functional relations like ‘holding up’, ‘obstructing’,

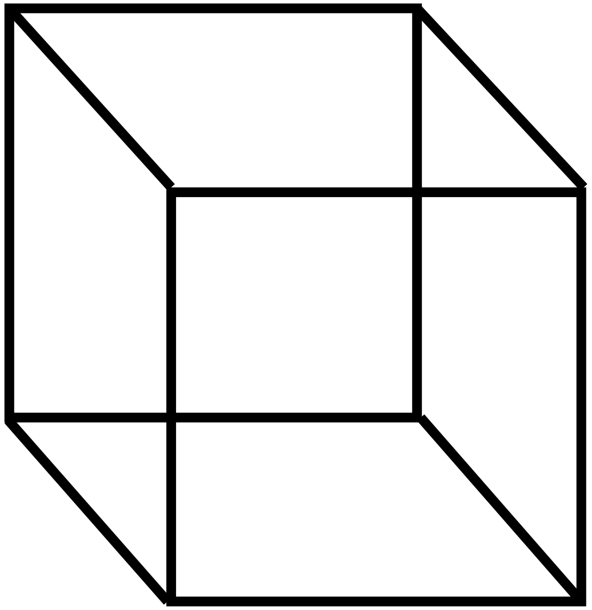
– seeing which way someone is facing,

– seeing how someone feels

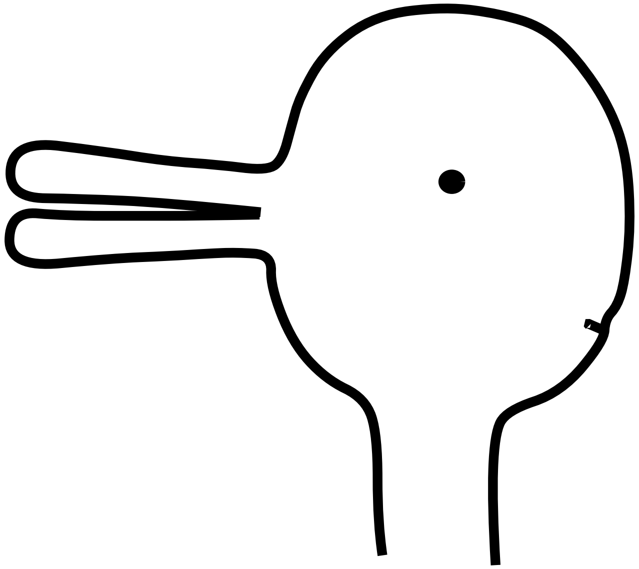
Examples: non geometric percepts

It is often thought that visual systems provide only information about geometrical properties and relationships of objects in the environment, plus surface properties like colour and texture; and also physical changes.

But some visually ambiguous figures suggest otherwise:



Necker cube



Duck-rabbit

What changes when the figures 'flip' ?

Necker Cube and Duck-rabbit

When the Necker cube figure flips, all the changes are geometric.

They can be described in terms of relative distance and orientation of edges, faces and vertices.

When the duck-rabbit flips the geometry does not change:

- The functional interpretation of the parts changes
- More subtle features change, attributable only to animate entities.

E.g. Which way is the animal looking?

These differences are visual, not simply inferential.
The examples occur in textbooks on vision, not reasoning

What does it mean to say that you "see the rabbit facing to the right".

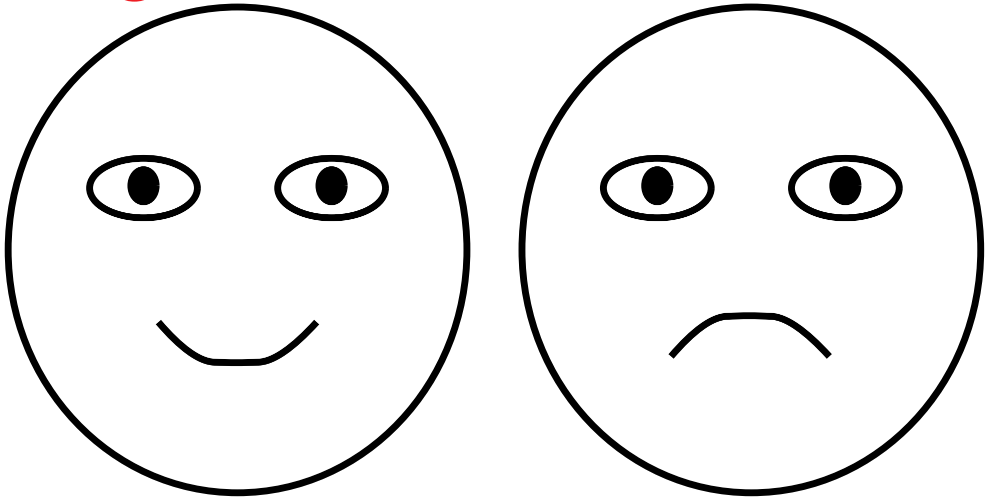
Perhaps it involves seeing the rabbit as a potential mover, more likely to move right than left.

Or seeing it as a potential perceiver, gaining information from the right.

What does categorising another animal as a perceiver involve?
How does it differ from categorising something as having a certain shape?

At the very least it involves using a meta-semantic ontology: an ontology with semantics that refers to objects that are themselves users of semantics, insofar as they refer to things.

Seeing Faces



Seeing facial expression as we do may just be a very old and simple process in which features of the face trigger reactions in a pattern-recognition device.

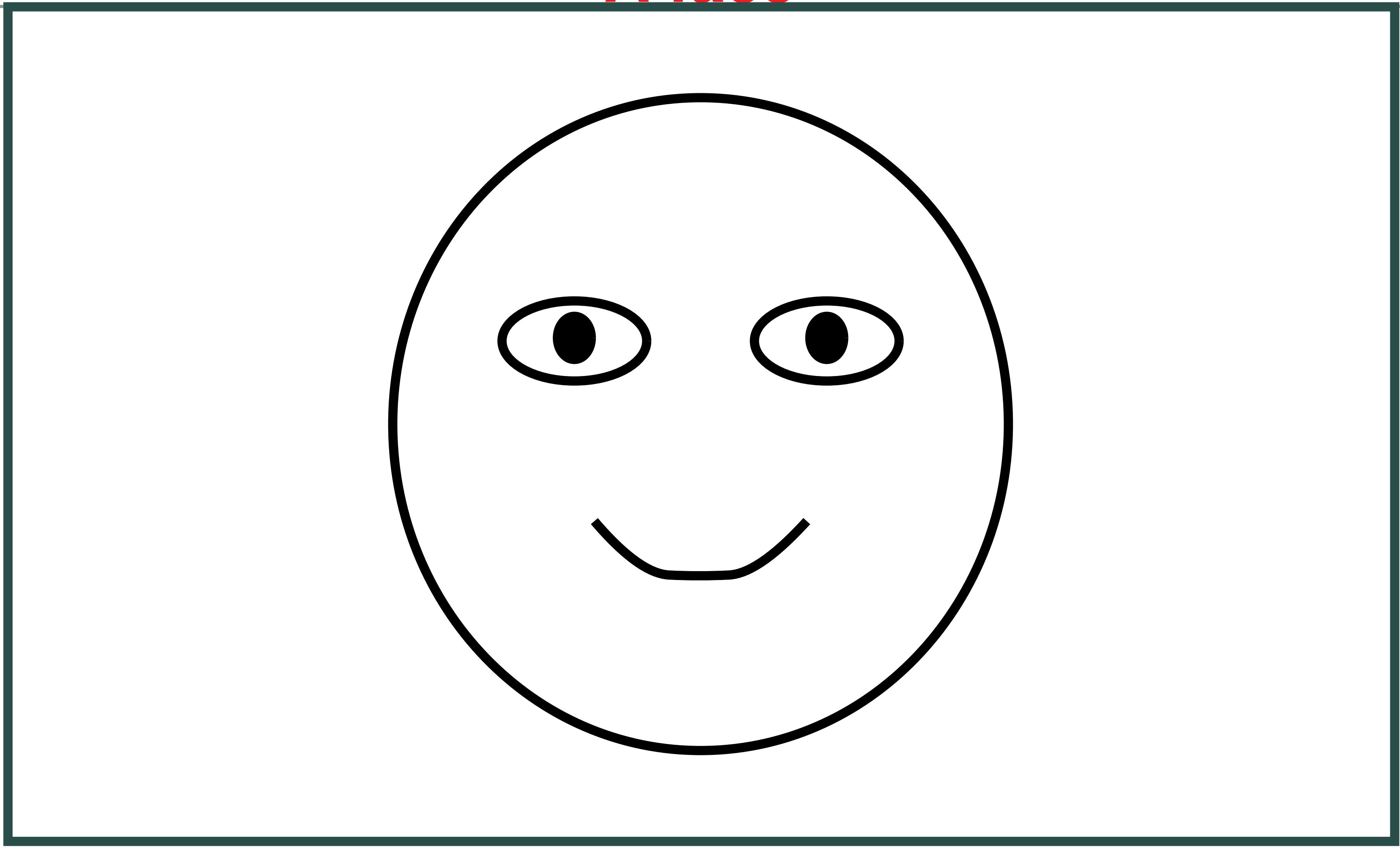
Or it may also involve deployment of sophisticated concepts that developed only through the evolution of meta-management. (Explained later)

For more on levels in perceptual mechanisms see the talk on visual reasoning and other talks here:
 http://www.cs.bham.ac.uk/˜axs/misc/talks/

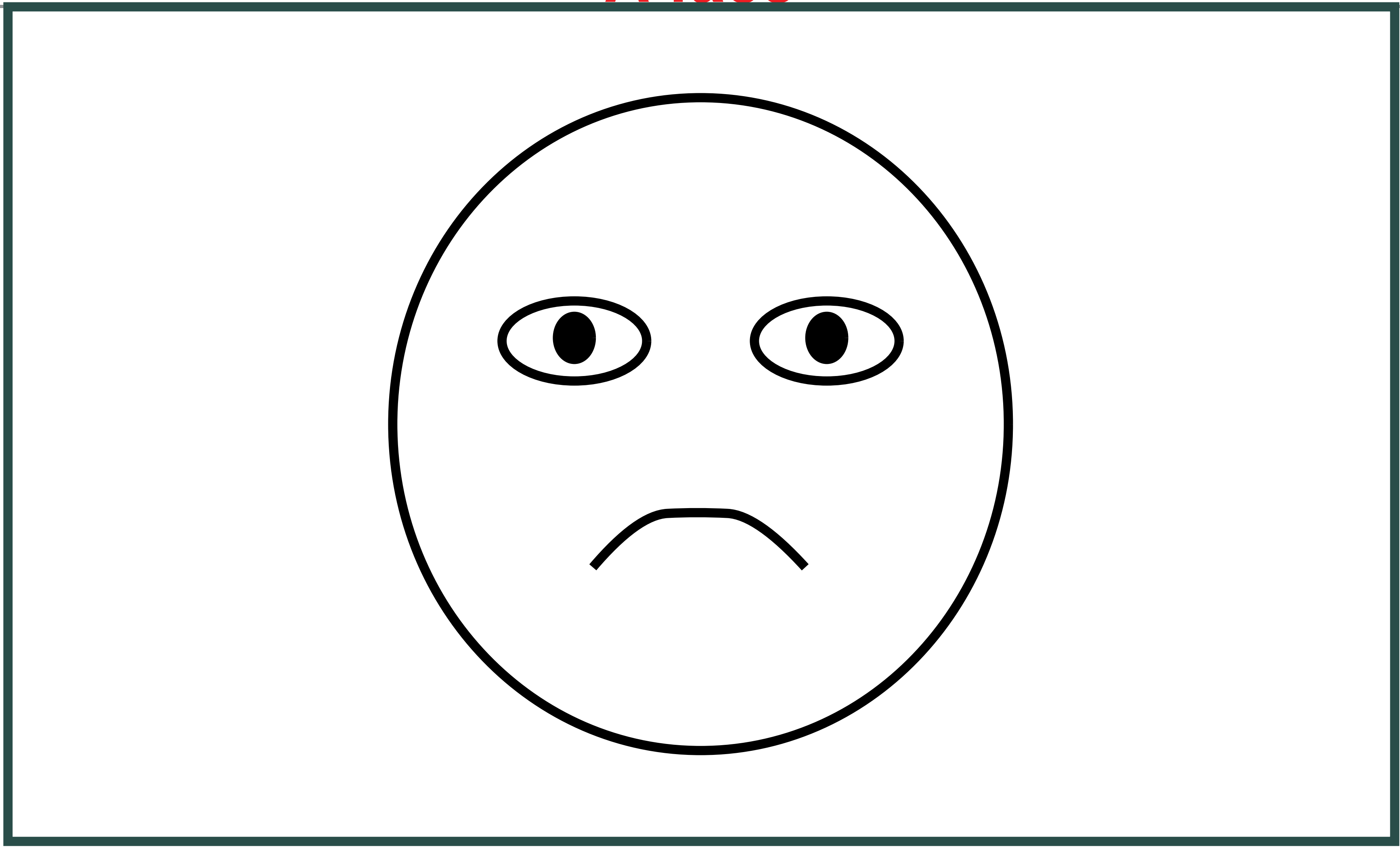
Some people see one pair of eyes as “looking happy” while the other pair “looks sad” or “looks angry”. (A non-geometrical context effect.)

Using the next two slides to flip rapidly between them may make this more evident.

A face



A face



Seeing mental states

What is involved in seeing an “expression”
e.g. happiness, sadness?

It is NOT just a matter of recognising and labelling a pattern.

Those visual categories are semantically linked to matters of importance to us as social animals,

just as the perception of geometric structure
is linked to our needs as agents in complex 3-D world
and our ability to act in that world.

Seeing how someone feels can affect what you should do next:
a non-geometric kind of affordance.

It seems to ‘colour’ the whole percept.

How can such a system be designed?

Can we build things that see happiness or sadness?

An appropriate architecture should explain the ability to have the sorts of percepts just discussed.

That ability requires at least some parts of the architecture to make use of an ontology that includes mental states – states that refer.

E.g. you are afraid of something.

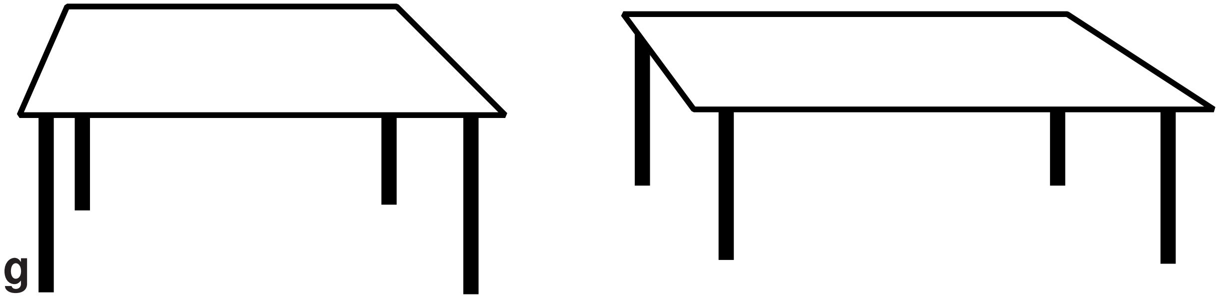
To see or think of or describe another as afraid, or wanting, or thinking requires a meta-semantic ontology: that refers to things that can refer, or more generally can process information.

(See H-Cogaff, later.)

Let's turn back to perception of physical objects and the affordances they provide.

Seeing tables

What sorts of affordances does a table provide?



– Obstruction
– Support
– Pulling, lifting, pushing,
  in various ways depending
  where you hold it and how.
– Easy availability of a collection of tools or papers, etc., in easy reach
– Social cohesion during meals
– Types of construction and repair methods
– ....

(See my 1996 'Actual possibilities' paper at the CogAff web site.)

Some of the affordances are conditional: e.g. you can pull the table if you (a) move closer and (b) grasp a leg or the edge.

How do we (and other animals) represent collections of possibilities and constraints on possibilities? How do we use our grasp of such possibilities and constraints to work out what to do?

Do we, or chimps, or crows, use modal logics?

Seeing possibilities and doing mathematics

Visual mathematical reasoning requires the ability to see not only structures but also

- **Possibilities for change**
- **Constraints on possibilities for change**
- **At various levels of abstraction**
- **E.g. metrical change, topological change, structural change**

The more complex a structure is the more possibilities for (small) changes it supports.

For more examples see talk 7 here:

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

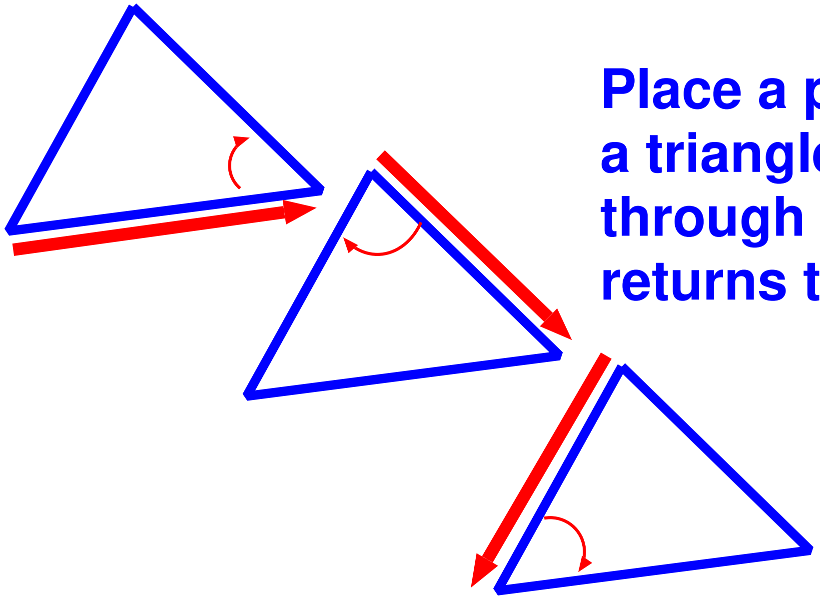
(Talk 7: Seminar slides on visual/spatial reasoning.)

Compare L.Wittgenstein's discussion of "seeing as" in his *Philosophical Investigations*, Part II, section (xi), 1953.

Seeing mathematical relations

There is a long history of people claiming that visual capabilities can be used for reasoning in everyday life and in mathematics.

E.g. how do you prove that the angles of a triangle add up to half a circle, i.e. 180 degrees.



Place a pencil along an edge of a triangle and rotate it in turn through the three angles until it returns to the original edge.

How much has it rotated?

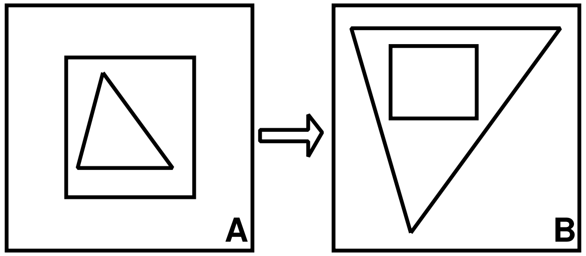
It is not necessary to use an actual triangle and pencil: the process can be visualised.

What difference does it make if you visualise *external* rotations of the pencil?

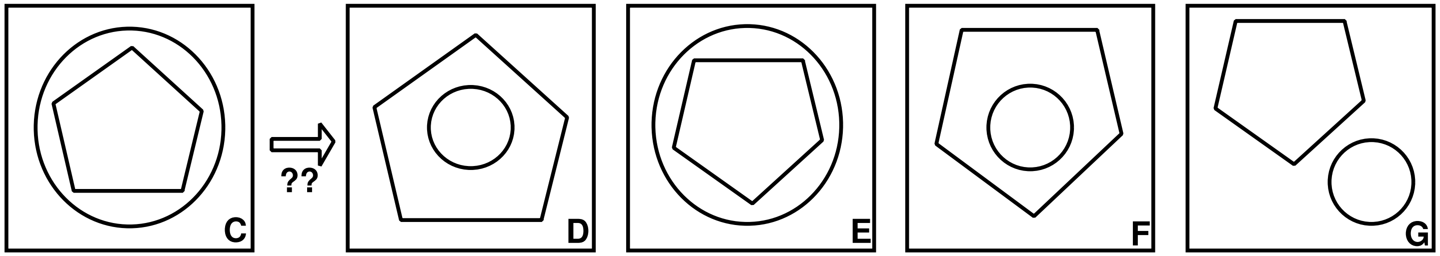
Seeing relationships between relationships

Standard intelligence test problems require one to see structures and to grasp not only relationships between parts of the structures, but also relationships between relationships (or, put another way, transformations of relationships).

How do we see those?



A is to B as
C is to which
of D, E, F, G?



An amazing program by T.G. Evans did this kind of thing in 1968:

‘A heuristic program to solve geometric analogy programs,’ in *Semantic Information Processing*, Ed., M.L. Minsky, MIT Press, pp. 271–353

(Could you program that sort of capability?)

More examples of visual reasoning

The ability to reason visually is part of everyday life

- How far should I lean over the table in order to be able to reach the salt cellar on the far side?
- Where should I stand in relation to the window in order to be able to see the left edge of the building opposite?
- How should I rotate that chair in order to get it through that door?
- Along which branch should I climb in order to be able to swing onto the next tree?
- Is the vase safely out of reach of that child?
- How should I cut these sheepskins in order to be able to assemble a jacket from the pieces?
- How can I design a mechanical loom, or a machine to make wind grind corn?

There are many activities that used visual reasoning long before the development of mathematics as we know it, but which may have used mechanisms that later made mathematical reasoning possible.

Even non-visual mathematical reasoning using algebraic and logical formulae requires us to be able to "see" structural relations in formulae, and to notice possibilities for syntactic transformations in those structures: more visual affordances.

Towards a taxonomy of uses of vision

I don't think anyone has attempted a systematic overview of the uses and capabilities of human and animal vision, including capabilities that are common to all and those that result from specialised training

The vast majority of visual affordances, and visual reasoning capabilities are not yet understood. (Contrast segmentation, recognition, distance estimation, tracking, ...)

Consider what it is to see a horizontal plane surface:

- Seeing it as having a uniform or changing texture or colour.
- Seeing it as separating the space above and below it.
- Seeing it as infinitely thin, or as indefinitely extendable.
- Seeing different parts as being at different distances from you.
- Seeing empty spaces as *possible locations* for a variety of shapes: lines, circles, pictures of faces, text, musical notation...
- Seeing parts of the surface as possible paths or trajectories.
- Seeing the possibility of a variety of processes in the plane: changing shape or texture, movement, pulsating objects, oscillations, etc.
- Seeing that the surface itself can move or rotate or bend in space.

Can we explain all this?

CONJECTURES:

- Animal visual architectures evolved several layers of analysis and interpretation.
- These operate concurrently, feeding information into different central layers which require different kinds of information represented in different ways (different affordances).
- Different aspects of human vision are related to differences in the functionality and sophistication of the central systems that they feed into.
- Likewise, there are likely to be different sorts of ‘top-down’ influences on visual processing, coming from different parts of the central architecture, with different requirements.
- In humans these include the ability to visualise what is not there and changes in what is there.
- Animals that have internal self-monitoring capabilities need conceptual apparatus for that task which can also be used in categorising mental states of other agents. (Meta-sematic capabilities.)

Evolution, the great philosopher/designer

In particular,

Evolution solved the “other minds problem” before anyone formulated it, by providing built-in apparatus for conceptualising mental states in others:

A requirement for

- prey species,
- predator species,
- social species.

We need an architectural framework in which to place all these diverse capabilities, as part of the design task.

Later we can modify the framework as we discover its limitations.

The framework should simultaneously help us understand the evolutionary process and the results of evolution.

How to reduce confusion and promote useful communication

Common terminology for discussing architectures would help.

We need a framework for thinking about the space of relevant architectures so that people taking design decisions can see:

(a) what the alternatives are

(b) for which purposes (niches) they are more or less appropriate.

(c) in which ways they are more or less appropriate for those purposes

This can help us with the following:

- Identifying the many uses/tasks of vision
  (different architectures, and different components within an architecture, need different kinds of visual information, or related information)
- Identifying the forms of representation useful for those tasks
- Taking the first steps towards explanatory theories and models.

There's no best or worst design – only trade-offs.

That's why such diverse biological solutions are all successful, in their own niches, e.g. microbes, insects, enormous varieties of plants and animals.

Beware of numerical evaluations (fitness functions): they lose information about what the strengths and weaknesses of the alternatives are.

So, if we want to understand the issues, evaluations should be primarily descriptive not numerical.

(Compare Consumer Association reports e.g. on lawn-mowers, or cars, or insurance providers.)

For more on this, see the papers on interacting trajectories in "design space" and "niche space" here:

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

e.g. the PPSN2000 paper.

Forms of representation in visual systems

Forms of representation studied so far for vision include

- 2-D rectangular arrays,
- concentric rings of receptive fields of varying size,
- weights or activations in neural nets,
- Fourier transforms,
- histograms and probability distributions,
- structural descriptions (parse trees),
- various symbolic representations of map structures,
- semantic nets,
- logical databases,
- control signals, ... and more

Biological vision probably uses forms of representation not yet thought of.

A hard problem: how to represent "affordances", and more generally information about possible changes and constraints on changes in a visible portion of the world.

See KR1996 paper on "Actual possibilities" at http://www.cs.bham.ac.uk/research/cogaff/

Generalising Gibson's notion of 'affordances'

Instead of thinking about visual 'affordances' for an organism, we think about the affordances for various components of an organism.

E.g. the following need different information from the environment, probably represented differently:

- posture control in two-legged walking
- control of visual saccades
- selection of routes
- building a shelter

The last task might include all the others!

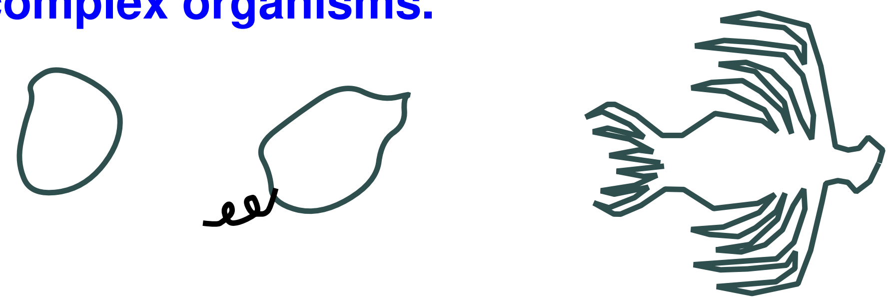
An architectural framework incorporating multiple mechanisms allows us to think about multiple visual pathways and multiple forms of learning and development.

We can then ask deeper questions about evolution: because we can formulate options with a deeper understanding of the space of designs and their trade-offs: e.g. trade-offs between species evolution and individual learning as means of acquiring information.

A biological perspective

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 to acquire, process, and use *information.*

The ability to act or to select requires information

E.g. information about

- density gradients of nutrients in the primaeval soup
- the presence of noxious entities
- where the gap is in a barrier
- precise locations of branches in a tree as you fly through
- how much of your nest you have built so far
- which part should be extended next
- where a potential mate is
- something that might eat you
- the grass on the other side of the hill
- what that thing over there is likely to do next
- how to achieve or avoid various states
- how you thought about that last problem
- whether your thinking is making progress

and much, much more... (has anyone attempted a taxonomy?)

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.

Things you can do with information

A partial analysis to illustrate the above:

- You can react immediately (it 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 try to derive new information from it (e.g. what caused this? what else is there? what might happen next? can I benefit from this?)
- You can store it for future use (and possibly modify it later)
- You can consider alternative next actions, or make plans
- 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

... using different forms of representation for different purposes.

The various kinds and uses of information-processing did not all evolve at the same time

Not all of them occur in all animals (microbes, insects, fishes, reptiles, birds, mammals, etc.)

A particular collection of sensory transducers (visual, auditory, tactile) can provide many different kinds of information at the same time, e.g. the text on the page, the window beyond the page, the state of the weather visible through the window, all in one visual field.

– Some information is very localised and simple (here's a dot, there's some motion).
– Other information may be far more holistic (e.g. recognising a scene as involving a forest glade).
– Some may be very abstract (the weather looks fine; it looks as if a fight is about to break out in that crowd).
– Some mechanisms involve only generally applicable knowledge about the geometry and topology of static and moving shapes.
– Others require specific knowledge about things that are relevant only in a particular part of the world, or a particular type of activity. E.g. seeing text, hunting fast moving prey, seeing geological formations, looking at exposed brains.

Diverse mechanisms of varying sophistication

Extracting information from the basic sensory data may require very diverse perceptual mechanisms with varying types of sophistication.

- Some information can be extracted very simply (using spatial or temporal local change detectors, or mechanisms for constructing histograms of features, such as colour, texture, optic flow).
- Other information may need *relationships* to be discovered between features, e.g. collinearity, lying on a circular arc, parallelism, closure, lying on the intersection of the continuations of two linear segments or two curved segments (where the continuations are also curved).
- Sometimes this requires *searching* for coherent interpretations.
- Some relationships hold only between abstract entities not the image data: e.g. two people seen to be *looking in the same direction*.
- Extracting some of the information requires matching with known models ("That's a triangle, a face, a tree").
- Some learning tasks require noticing new repeated structures within the information structures (e.g. noticing repeated occurrence of polygons with circles at two adjacent corners).

Virtual vs physical machines

In computer science, software engineering and AI we have learnt the importance of virtual machines, e.g. the Lisp, Prolog, Java virtual machines, chess virtual machines, neural net simulations, etc.

Mechanisms that operate on complex information structures are typically virtual machines (parsers, structure matchers, network constructors, search engines, planners, interpreters, etc.) rather than physical machines, though virtual machines are *implemented* in physical machines.

This implies that if we are to explore the full range of architectures for intelligent systems, including architectures for visual systems, we need to be familiar with a wide range of techniques for constructing virtual machines of various sorts.

This has implications for the sorts of education that should be provided for broad-minded AI students.

For more on the relation between virtual machines and physical machines (a hard philosophical problem) see the slides for my IJCAI tutorial with Matthias Scheutz:

http://www.cs.bham.ac.uk/~axs/ijcai01/

Temporal and causal differences in virtual machines for vision

Some perceptual information is used "online",

e.g.

- Posture control
- Control of a hand moving to pick up a pencil, or a pin, or to pick a berry in a thorny plant.
- Use of vision when parking your car
- Reading text aloud, or sight-reading a musical score as you play.

Some is stored for future use, in various modes, e.g.

- Recognising the person who punched you a week ago
- Remembering where you put a pencil
- Learning a new discrimination (e.g. learning to distinguish a pair of identical twins)
- Formulating generalisations
  (Xs are found inside Ys, Doing A to X, causes X to do B)
- Storing a plan that is found to be useful.
- Many perceptual-motor skills produced by training

Often online control can use *continuous* variation, whereas much stored information concerns *discrete* categories and relationships.

The evolution of information processing architectures and mechanisms

Evolution "discovered" and used many things long before human engineers and scientists asked the questions: long before they even existed.

Paleontology shows the development of physiology and provides some weak evidence about behavioural capabilities.

But there is very little direct evidence regarding previous forms of information processing: virtual machines leave no fossils.

Archaeologists speculate wildly and (in my view) irresponsibly.

We can be more disciplined!

The variety of forms of information processing now found in nature gives many clues, and we can test theories in working models.

Some of the forms are evolutionarily very old. Others relatively new. (E.g. the ability to learn to read, design machinery, or do mathematics.)

WE NEED TO LEARN HOW TO ASK GOOD (DEEP) QUESTIONS.

Different information processing architectures

The different tasks require different kinds of mechanisms, often operating on different forms of representation and different forms of long and short term storage.

Sometimes they require different sub-mechanisms working together (perceiving, learning, using prior knowledge, deciding what to do, constructing plans, executing plans, etc.)

But there must always be an ARCHITECTURE combining all the mechanisms and processes they produce.

Some of the more sophisticated mechanisms and architectures evolved only relatively recently, and are in very few species (e.g. deliberative capabilities – see below)

We need to understand how they differ from, how they are built on, and how they interact with the much older, more wide-spread mechanisms.

The same organism, e.g. a human being, may include both very old and very new mechanisms, in many sub-systems.

Some differences are very subtle

Physiological and other similarities between visual systems of different mammals, e.g. lions and sheep, may mislead us.

There may be subtle, unobvious, but very important differences, e.g. where one organism has a mostly genetically determined information processing architecture, whereas another builds much of its architecture using a boot-strapping process after birth.

The results may be very different in the capabilities they support.

E.g. a grazing mammal and a hunting mammal have very different visual requirements.

Likewise compare birds that just peck grains on the ground and nest at ground-level (chickens) with hunting birds that build tree-top nests (magpies).

Biologists distinguish:

– precocial species (e.g. deer, chickens)

– altricial species (e.g. lions, eagles).

Precocial species are born or hatched more physiologically developed and more behaviourally competent: why?

A clue: look at the different (adult) niches

We need to analyse and contrast the visual requirements of adults:

– grazing mammals (e.g. deer)
– hunting mammals (e.g. lions).

How do their visual tasks differ?

What are the implications of the requirement to be able to stalk, to chase, and then jump and bite the neck of a fast moving animal?

Will a deer and a lion see the same things if they look at the same terrain?

Compare the grasp of spatial structure and motion required for use of a hand with opposing thumb, in picking berries or moving small insects from tree branch to mouth.

Contrast that with the visual requirements of a bird that pecks at such berries or insects.

Contrast using *your own* hand to pick berries with watching *how another person* does it: the tasks are different in subtle ways.

Conjecture: bootstrapping in altricial species

In precocial species evolution can pre-program the visual capabilities required, and they are available to the young almost immediately.

In altricial species (e.g. hunting mammals, nest building birds, apes and monkeys) the activities of adults require a far more sophisticated visual grasp of structure and motion (and links to tactile perception).

Specifications for mechanisms that have all the latter information may be too complex to encode in genes.

But it may be possible to encode a bootstrapping system that causes the required mechanisms to be developed while the architecture grows, using a variety of exploratory actions including infant play.

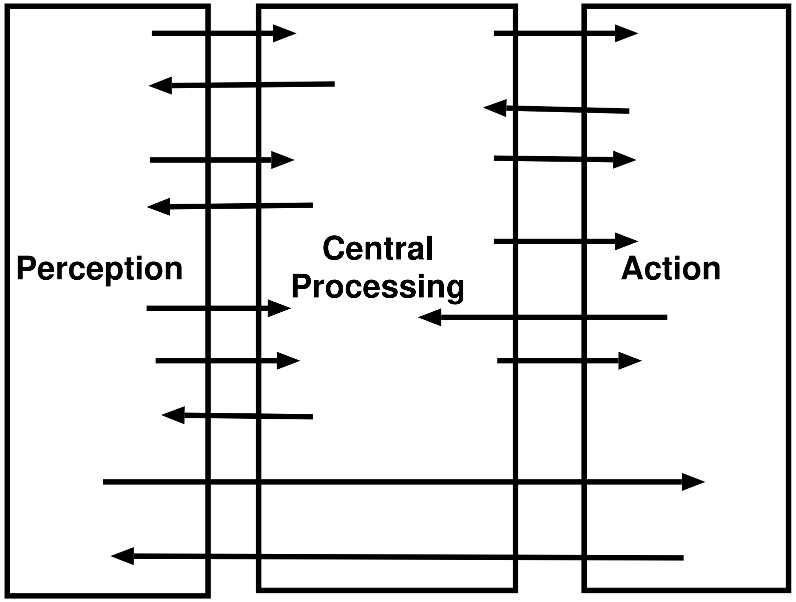
HOW IS THAT ACHIEVED ????

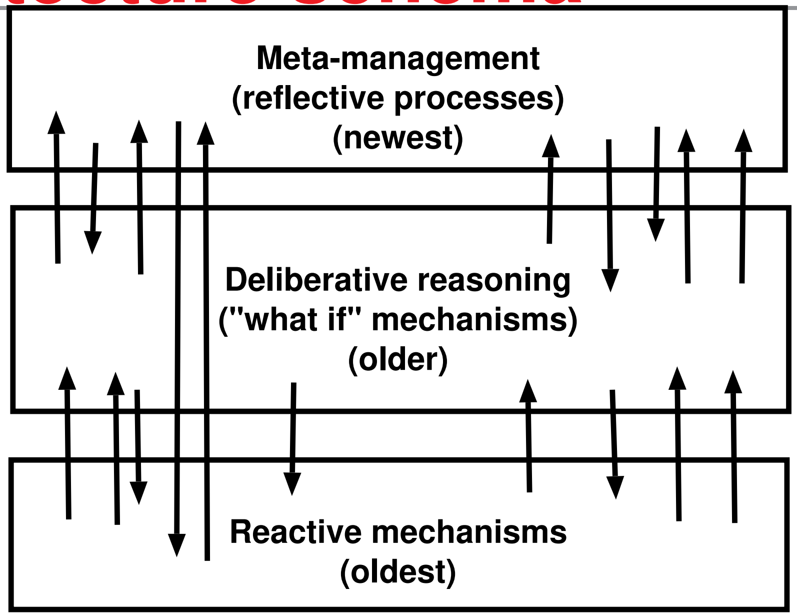
Could it all be just calibration, e.g. using play, etc., to specify quantitative parameters, within a fixed architecture?

I doubt it, but that's an open question.

Human learning capabilities, e.g. learning to speak or read, seem to arise from more general bootstrapping mechanisms.

Towards an architecture schema





Two coarse divisions within information processing architectures – 'towers' and 'layers':

- **(a) Nilsson's (1998) "triple tower" model**
- **(b) Layered architectures: e.g. reactive, deliberative and meta-management layers.**

(a) and (b) express different (orthogonal) functional divisions.

These divisions can be combined, as follows ....

Superimposing the divisions: The COGAFF Schema

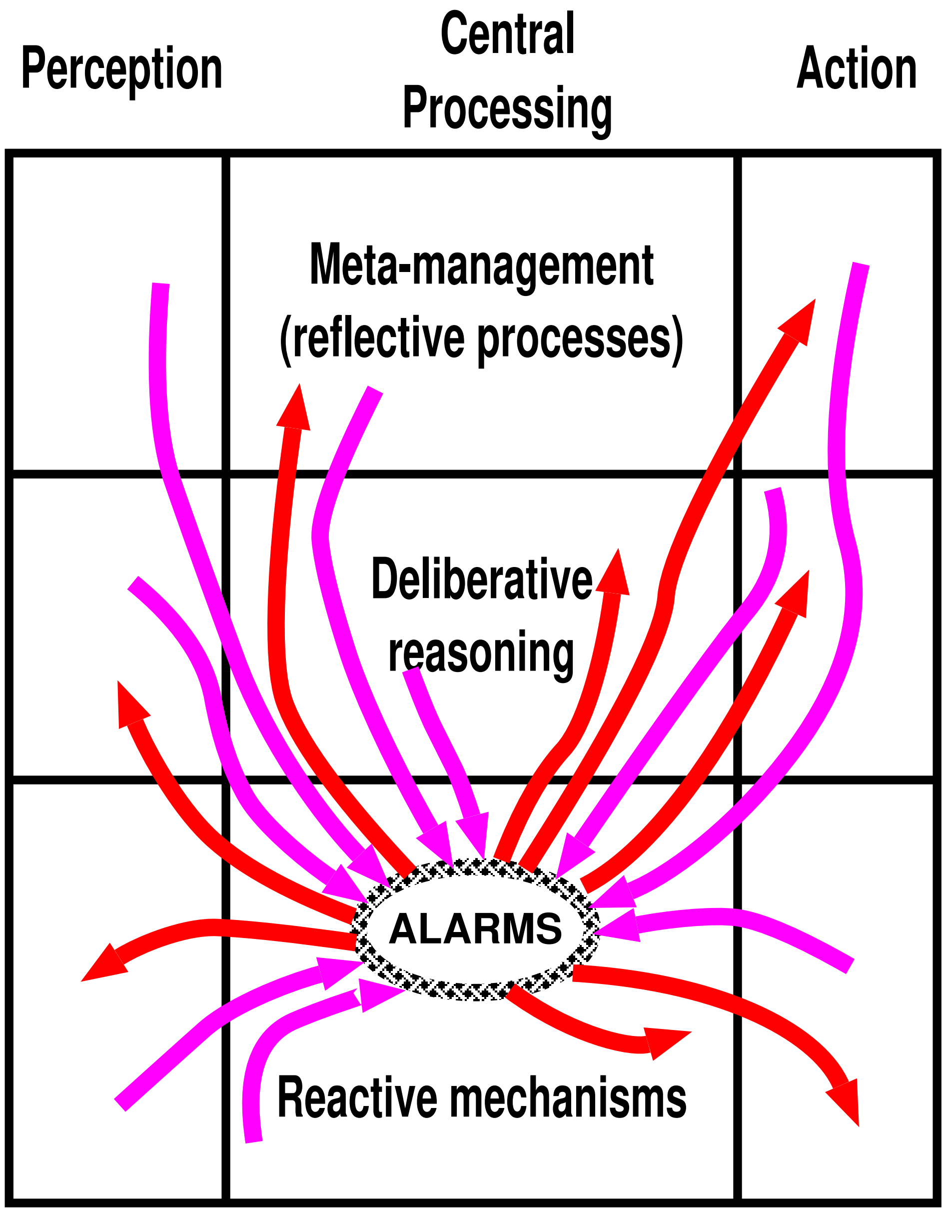
Perception	Central Processing	Action
	Meta-management (reflective processes) (newest)	
	Deliberative reasoning ("what if" mechanisms) (older)	
	Reactive mechanisms (oldest)	

Boxes indicate possible functional roles for mechanisms: only some possible information flow routes are shown (cycles are possible within boxes, but not shown).

COGAFF extended – with “alarm mechanisms”

Alarm mechanisms deal with the need for rapid reactions using fast pattern recognition based on information from many sources, internal and external.

An alarm mechanism is likely to be fast and stupid, i.e. error-prone, though it may be trainable.



Characterising the layers

The differences between the layers are complex and subtle.

Some of the differences are discussed in other slide presentations here
http://www.cs.bham.ac.uk/~axs/misc/talks/

Further discussion is in the papers in the Cogaff directory
http://www.cs.bham.ac.uk/research/cogaff/

It may turn out that there are better ways of dividing up levels of functionality, or that more sub-divisions should be made – e.g. between analog and discrete reactive mechanisms, between reactive mechanisms with and without chained internal responses, between deliberative mechanisms with and without various kinds of learning, or with various kinds of formalisms, and between many sorts of specialised “alarm” mechanisms.

The COGAFF schema is still a draft, likely to evolve

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, there can be multi-window action or peephole action.

Architectural change in an individual

Learning can introduce new architectural components, e.g. the ability to read music, the ability to write programs.

Development of skill (speed and fluency) through practice can introduce new connections between modules, e.g. links from higher-level perceptual layers to specialist reactive modules.

For instance, learning to read fluently, or developing sophisticated athletic skills.

Highly trained skills can introduce new “layer-crossing” pathways, e.g. visual pathways: rapid recognition of a category originally developed for deliberation can, after training, trigger fast reactions.

Cogaff is a schema not an architecture: a sort of 'grammar' for architectures

Different organisms, different artificial systems, may have

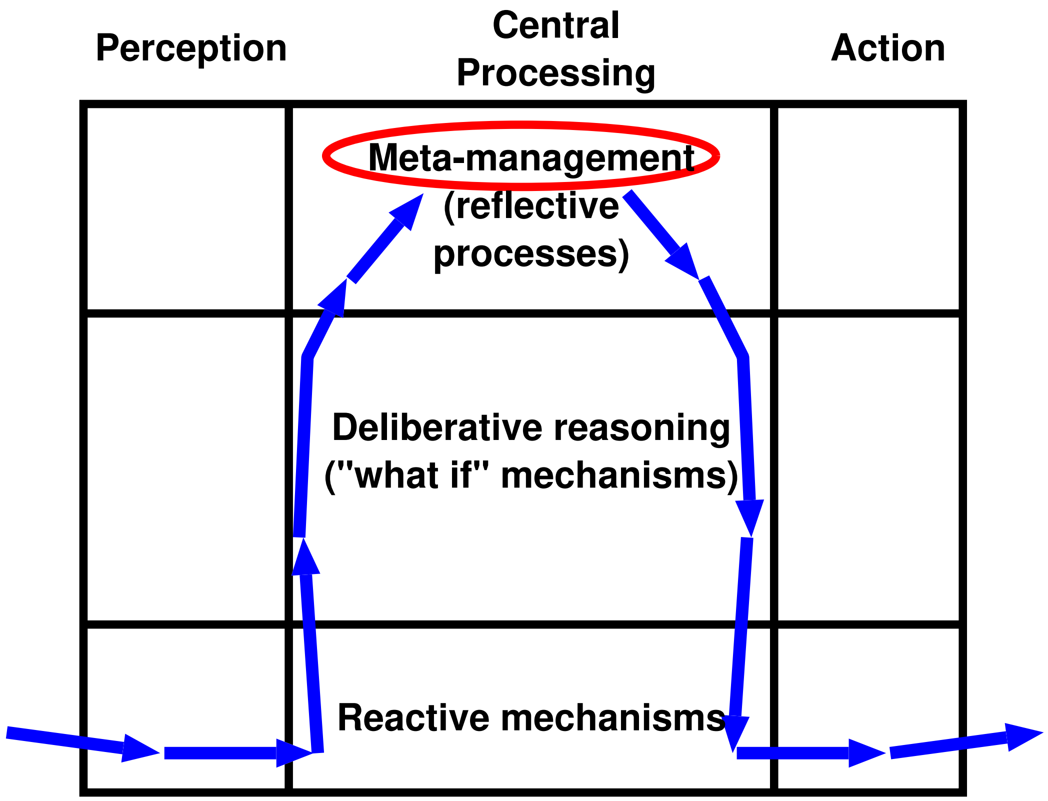
- different components of the schema
- different components in the boxes
- different connections between components

E.g. some animals, and some robots have only the reactive layer (e.g. insects, microbes).

The reactive layer can include mechanisms of varying degrees and types of sophistication, some analog, some digital, with varying amounts of concurrency.

Other layers can also differ between species.

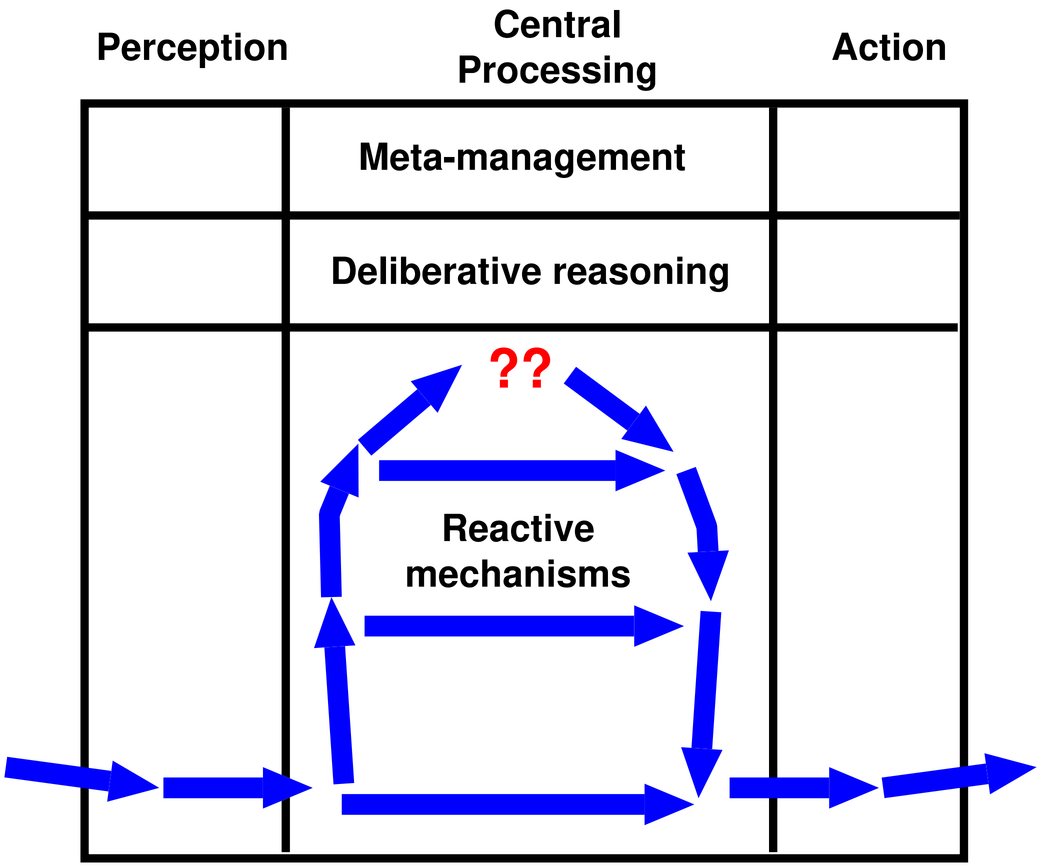
An example sub-category: Omega architectures



This is just a pipeline, with “peephole” perception and action, as opposed to “multi-window” perception and action.

E.g. Cooper and Shallice: Contention scheduling, Albus 1981.

Another sub-category: Subsumption architectures (R.Brooks)



This could be useful for certain relatively primitive sorts of organisms and robots. (E.g. Insects, fish, crabs?)

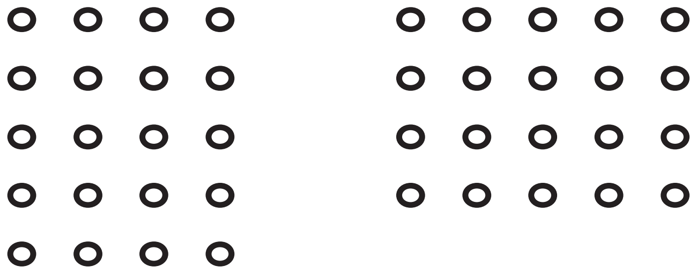
There are discontinuities in design space

E.g. in humans the deliberative and meta-management layers appear to have unique mechanisms and forms of representation, not found in other animals.

Can a chimp (or bonobo) think about the relation between mind and body?

Or learn about predicate calculus and modal logic?

Or see the structural correspondence between these two?

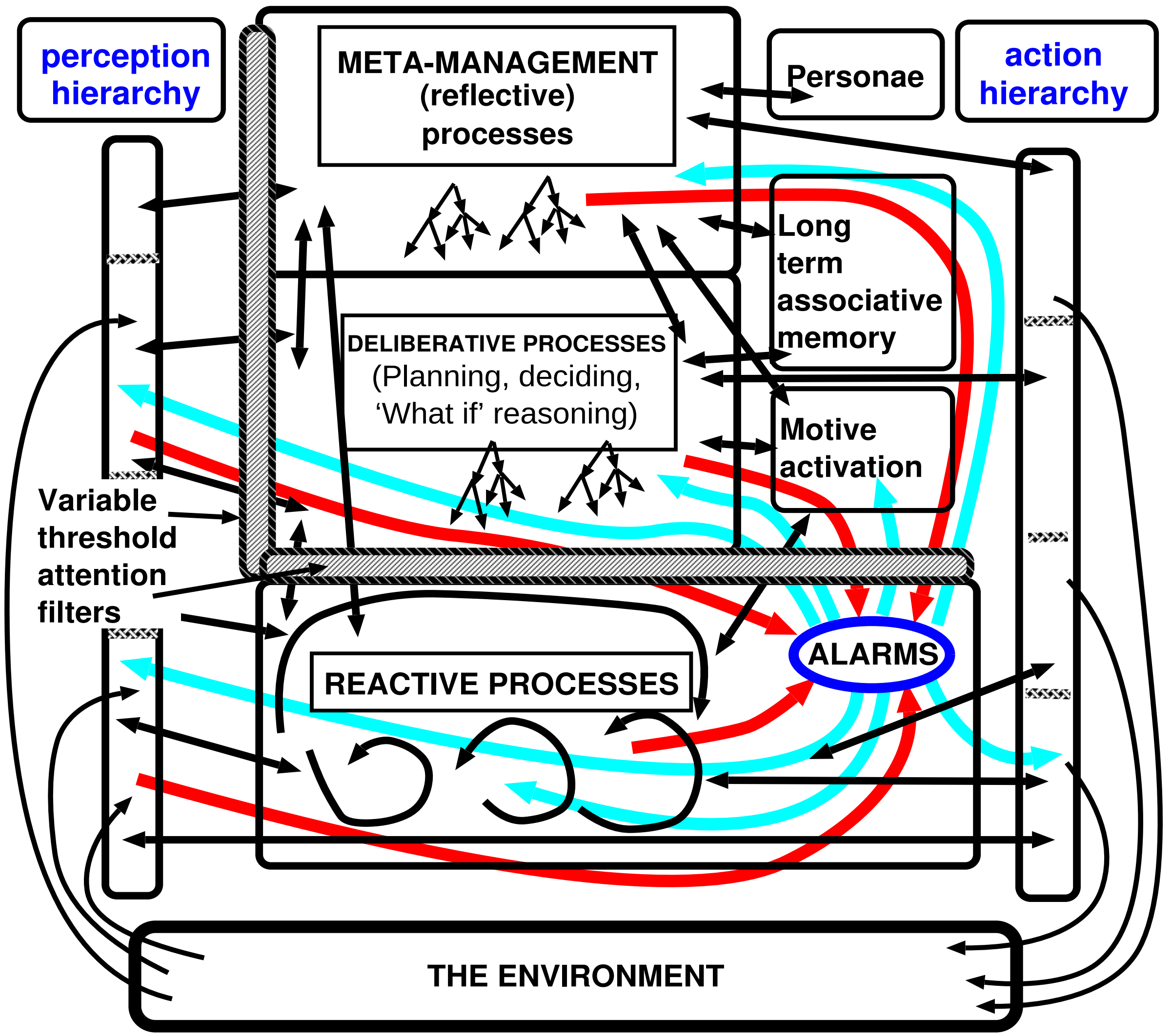


(I don’t know the answers - apes can do amazing things.)

It is not clear that all discontinuities are results of sequences of very small changes. Darwinian evolution might sometimes (rarely!) produce large useful changes. (DNA is a discrete structure.)

The “Human-like” sub-schema H-Cogaff

Our conjectured architecture for human-like systems:



The “Human-like” sub-schema H-Cogaff

The reactive, deliberative and meta-management layers evolved at different times, requiring discontinuous changes in the design, and providing significantly new capabilities.

An attention filter with dynamically varying threshold may be used to protect resource-limited higher level functions.
(Luc Beaudoin’s PhD thesis 1994)

Some aspects of the alarm system apparently correspond to the brain’s limbic system.

Frontal lobes apparently implement some meta-management functions.

See the Cogaff papers:
http://www.cs.bham.ac.uk/research/cogaff/

Some implications

Within this framework we can explain

- research findings on different visual pathways (and predict more)
- blindsight (damage to some meta-management access routes prevents self-knowledge about some visual processes)
- varieties of emotions (at least three distinct types related to the three layers: primary, secondary and tertiary emotions)
- many varieties of learning and development
- the discovery by philosophers of ‘qualia’
- some of the evolutionary trade-offs in developing these systems (Higher levels can be very expensive, and require a food pyramid)

and probably much more

Warning to experimenters

Of the many forms of concurrent perception in different parts of the architecture, we are *aware of* only those aspects accessible to the meta-management processes.

So we cannot report verbally on, or otherwise voluntarily indicate the presence of, the others, including:

- **some of the perceptions in the reactive sub-system which influence reactive behaviours**
- **some of the intermediate stages in visual processing which produce percepts that meta-management can access: e.g. we may be unaware of intermediate stages in producing percepts of objects in the environment, even where we are aware of results of *later* stages**

So we cannot assume that asking subjects questions in experiments, or getting them to press buttons or turn dials to indicate what they see is a reliable way to find out everything they can see.

I.e. this theory implies that there are forms of "blindsight" in normal humans: it is not just a product of brain damage.

But much of this is still far too vague

There is a huge amount of work still waiting to be done

Including working out in great detail:

- what sorts of visual capabilities are possible (in humans and other animals)
- and how they relate to niche features (PPSN2000 paper),
- and then investigating ways of explaining and implementing them.

This will very likely require us to discover:

- new forms of representation,
- new information processing mechanisms for manipulating them,
- new architectures to incorporate and make use of those mechanisms.
- new characterisations of what such such architectures can use vision for (e.g. seeing *possibilities* and *impossibilities*)

What I am not saying

I am not saying that everyone should drop what they are doing and join this ambitious integrative research programme.

It would be nice to have a few more people thinking about it from time to time, however.

Perhaps we can revive the endangered species.

Many thanks to the conference organisers (of BMVC'01)
for giving me the opportunity to try

And now the European Commission has give us the opportunity to work on a collaborative project to test these ideas.

See the CoSy project web site:

http://www.cs.bham.ac.uk/research/projects/cosy/PlayMate-start.html

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

For more on all this see

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

http://www.cs.bham.ac.uk/~axs/misc/talks
(slides for several talks, including this one)

http://www.cs.bham.ac.uk/research/poplog/freepoplog.html
(software tools for exploring hybrid architectures)

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0505
(A partly new theory of vision as process simulation)

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0506
(On children leaning to understand causation).

This is yet another linux-only presentation