A Multi-picture Challenge for Theories of Vision

Aaron Sloman

School of Computer Science, University of Birmingham

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

These PDF slides are in my 'talks' directory:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk88
and will be added to slideshare in flash format:
http://www.slideshare.net/asloman/presentations/multipic-challenge

NOTE: some of these ideas were anticipated by the psychologist Julian Hochberg and developed in collaboration with Mary Peterson (Peterson, 2007),

Last Changed (June 2, 2013): liable to be updated.

The Problem

- Human researchers have only very recently begun to understand the variety of possible information processing systems.
- In contrast, for millions of years longer than we have been thinking about the problem, evolution has been exploring myriad designs.
- Those designs vary enormously both in their functionality and also in the mechanisms used to achieve that functionality

  including, in some cases, their ability to monitor and influence some of their own information processing (not all).
- Most people investigating natural information processing systems assume that we know more or less what they do, and the problem is to explain how they do it.
- But perhaps we know only a very restricted subset of what they do, and the main initial problem is to identify exactly what needs to be explained.
- Piecemeal approaches can lead to false explanations: working models of partial functionality (especially in artificial experimental situations) may be incapable of being extended to explain the rest – modules that "scale up" may not "scale out".
- My concerns are not primarily with neural mechanisms: I think it is important to get clear what sort of virtual machines are implemented on them – especially what their visual functions are.
- This is one of several papers and presentations probing functions of vision: speculations about suitable mechanisms to meet the requirements are in my other presentations.

Constraints on mechanisms
The problem of speed

One of the amazing facts about human vision is how fast a normal adult visual system can respond to a complex optic array with rich 2-D structure representing complex 3-D structures and processes, e.g. turning a corner in a large and unfamiliar town.

The pictures that follow present a sequence of unrelated scenes.

Try to view the pictures at a rate of one per second or less: i.e. keep your finger on the "next-page" button of your preferred PDF viewer. (Mine is 'xpdf' on Linux.)

Some questions about the pictures are asked at the end.

Please write down your answers (briefly) then go back and check the pictures.

I would be grateful if you would email the results to me. (A.Sloman@cs.bham.ac.uk)
Suggestions for improving or extending the experiment are also welcome.

Pictures are coming

What do you see and how fast do you see it?

Pictures selected from Jonathan Sloman's web site, presented here with his permission.

The problem of speed

(1)

The problem of speed

(2)

The problem of speed

(3)

The problem of speed

(4)

The problem of speed

(5)

The problem of speed

(6)

Some questions

Without looking back at the pictures, try to answer the following

1. What animals did you see?
2. What was in the last picture?
3. Approximately how many goats did you see? Three? Twenty? Sixty? Some other number?
4. What were they doing?
5. What was in the picture taken at dusk?
6. Did you see any windows?
7. Did you see a uniformed official?
8. Did you see any lamp posts? Approximately how many? Two? Ten? Twenty? Sixty? Hundreds?
9. Did you see a bridge? What else was in that picture?
10. Did you see sunshine on anything?
11. What sort of expression was on the face in the last picture?
12. Did anything have hands on the floor?
13. Did anything have horns? What?

Now look back through the pictures and note what you got right, what you got wrong, what you missed completely. Any other observations? (Email: A.Sloman@cs.bham.ac.uk)
(Don't expect to be able to answer all or most of them: The problem is how people can answer ANY of them.)

More pictures follow, with more questions at the end.

The problem of speed

(7)

The problem of speed

(8)

The problem of speed

(9)

The problem of speed

(10)

The problem of speed

(11)

The problem of speed

(12)

The problem of speed

(13)

Some more questions

Without looking back at the pictures, try to answer the following

1. What animals did you see?
2. What was in the last picture?
3. Were any views taken looking in a non-horizontal direction?
4. Did you see any windows?
5. Was anything reflected in them?
6. Was anything behind bars?
7. Did anything mention Brixton?
8. Did you see a curved building? Did it have windows?
9. What sort of expression was on the face in the last picture?
10. Which way did the two black arrows point?
11. Did you see anything from a science fiction series?
12. Did you see any words on the floor? What did they say? What colour were the letters?
13. Did anything have hands on the floor?
14. Did you see an egg? Do you remember anything about it?
15. In the picture with several pairs of legs were they legs of adults or of children?
16. What was white and broken?

Now look back through the pictures and note what you got right, what you got wrong, what you missed completely. Any other observations? (Email: A.Sloman@cs.bham.ac.uk)
(Don't expect to be able to answer all or most of them: The problem is how people can answer ANY of them.)

Note added 20 Mar 2012: I've learnt from Kim Shapiro that experiments like these were done long ago by Mary C Potter (MIT) `http://mollylab-1.mit.edu/lab/publications.html#1`

Beyond these Informal Experiments

Could these demonstrations and questions be turned into a more precise experiment?

Would anything be learnt by getting people to do this in a brain scanner?

Has something like it been done already ?

Has anyone proposed a model or identified mechanisms that explain how these competences work?

It should be a model that has so much content and is so precise that it could be used by a suitably qualified engineer as the design for a machine that could be built to demonstrate what the model explains.

Some sketchy suggestions follow.

Towards a design for a working model

The phenomena above and many others described in the papers and presentations below have led me to the conclusion that we need to think of human brains as running virtual machines of the second sort depicted.

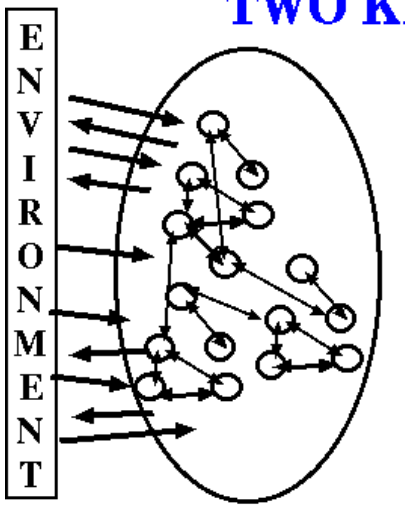
Some subsystems change rapidly and continuously, especially those closely coupled with the environment, while others change discretely, and sometimes much more slowly, especially those concerned with theorising, reasoning, planning, remembering, predicting and explaining, or even free-wheeling daydreaming, wondering, etc.

At any time many subsystems are dormant, but can be activated very rapidly by constraint-propagation mechanisms.

The more complex versions grow themselves after birth, under the influence of interactions with the environment (e.g. a human information-processing architecture).

[See papers by Chappell and Sloman]

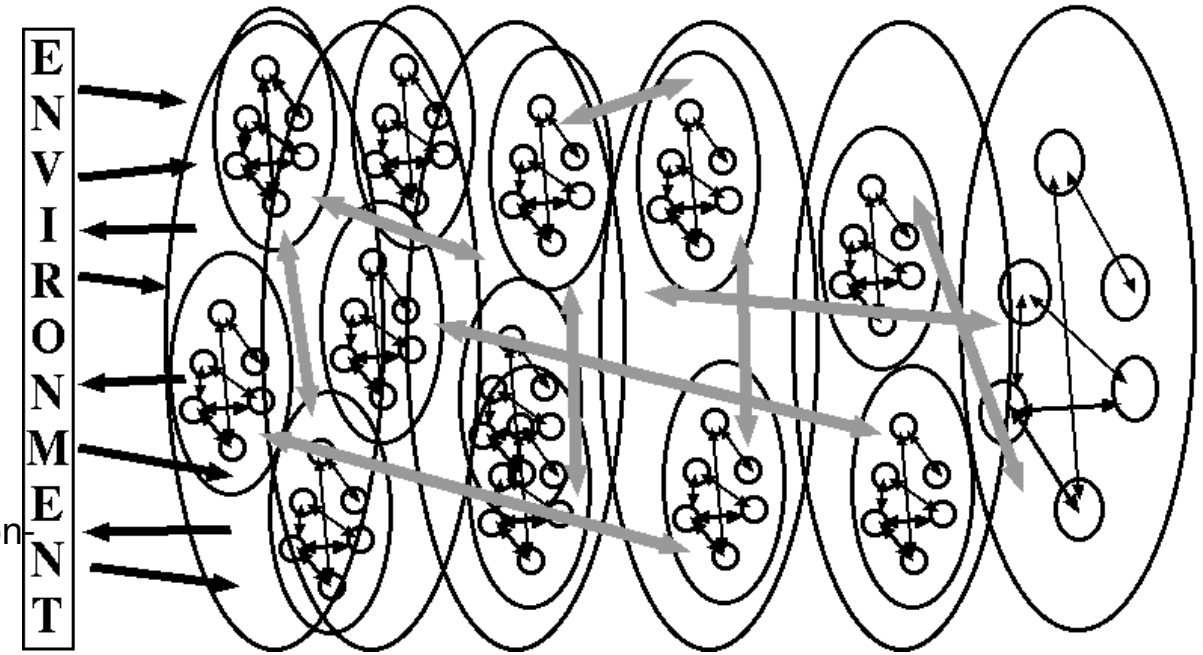
TWO KINDS OF DYNAMICAL SYSTEM



One kind, on the left, is closely coupled with the environment through sensors and effectors.

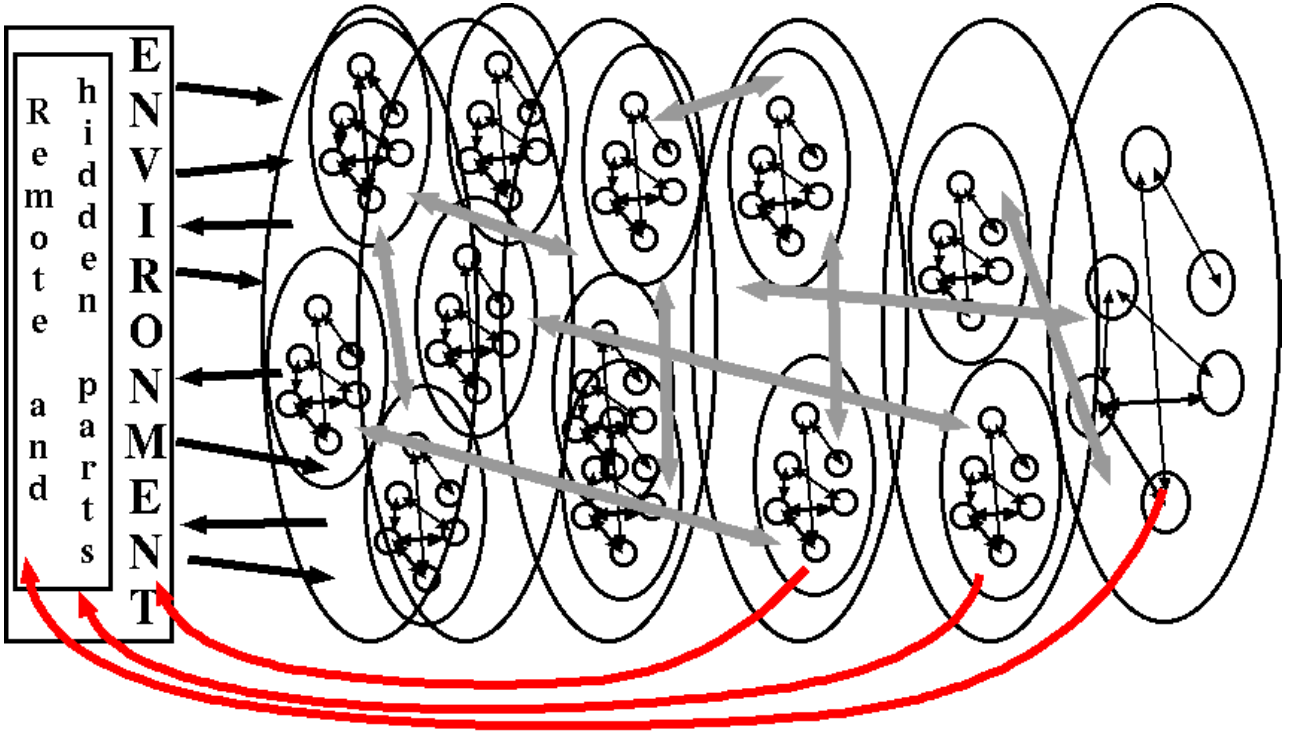
The other kind, below, has many levels of abstraction and decoupling from the environment, including theorising, reasoning remembering, planning, predicting subsystems.

Evolution produced both kinds and many intermediate kinds.



A computer model of a variant of this idea was reported in 1978 in *The computer revolution in philosophy, Ch 9*
`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/chap9.html`

Somatic and exosomatic semantic contents

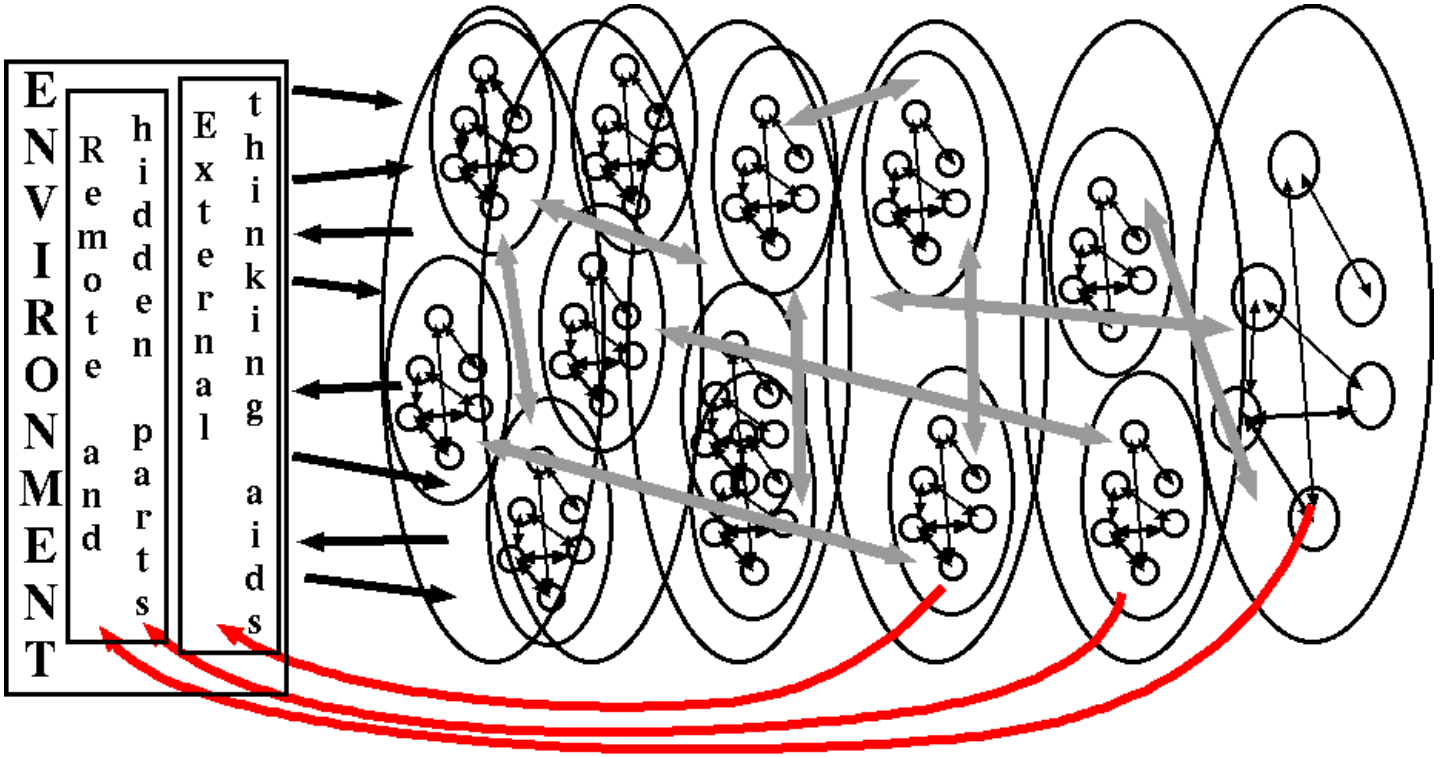


Some layers have states and processes that are closely coupled with the environment through sensors and effectors, so that all changes in those layers are closely related to physical changes at the interface: The semantic contents in those interface layers are “somatic”, referring to patterns, processes, conditional dependencies and invariants in the input and output signals.

Other subsystems, operating on different time-scales, with their own (often discrete) dynamics, can refer to more remote parts of the environment, e.g. internals of perceived objects, past and future events, and places, objects and processes existing beyond the current reach of sensors, or possibly undetectable using sensors alone: These can use“exosomatic” semantic contents, as indicated by red lines showing reference to remote, unsensed entities and happenings (including past and future and hypothetical events and situations).

For more on this see these talks `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/`
and papers by Chappell & Sloman in `http://www.cs.bham.ac.uk/research/projects/cosy/papers/`

Thinking with external objects



As every mathematician knows, humans sometimes have thought contents and thinking processes that are too complex to handle without external aids, such as diagrams or equations written on paper, blackboard, or (in recent years) a screen.

I pointed out the importance of this in (Sloman, 1971), claiming that the cognitive role of a diagram on paper used when reasoning, e.g. doing geometry, or thinking about how a machine works, could be functionally very similar to the role of an imagined diagram.

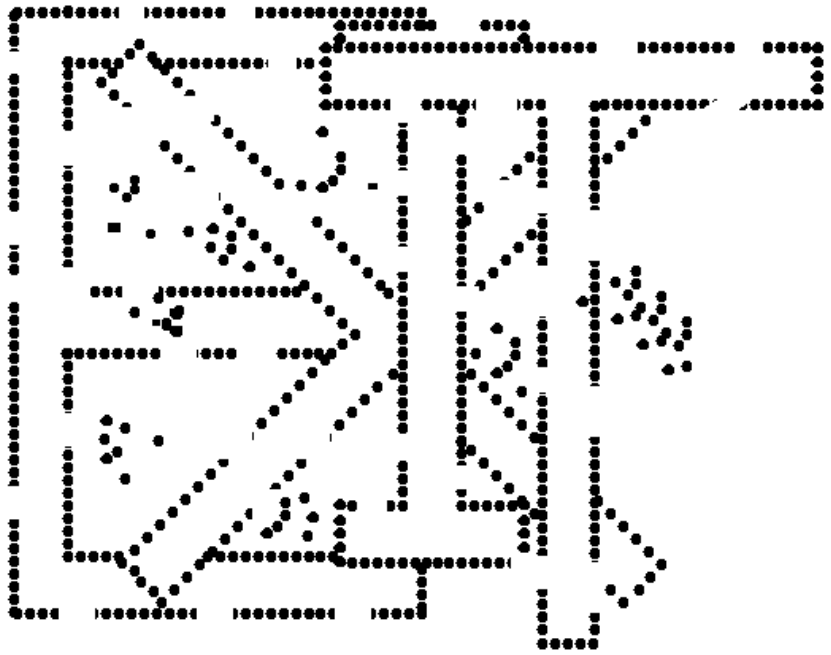
If the mechanisms of the machine are visible, they can play a similar role.

Chapters 6 and 7 of (Sloman, 1978) extended these ideas, including blurring the distinction between the environment and its inhabitants.

The Popeye Project

Chapter 9 of Sloman (1978) included a summary description of the POPEYE computer program developed (with David Owen, Geoffrey Hinton and Frank O'Gorman) at Sussex University.

The program could be presented with pictures of the sort shown, where a collection of opaque "laminas" in the form of capital letters could be shown in a binary array, with varying amounts of positive and negative noise and varying difficulty in interpreting the image caused by occlusion of some letters by others.



By operating in parallel in a number of different domains defined by different objects, properties and relationships, and using a combination of top-down, bottom up, middle-out and sideways information flow (constraint propagation), the program was able to find familiar words that might otherwise be very hard to see, and often it would recognise the word before completing processing at lower levels.

The program's speed and accuracy in reaching a first global hypothesis depended on the amount of noise and clutter in the image, and on how many words the program knew with common partial sequences of letters.

In other words, it exhibited "graceful degradation".

The next slide illustrates the domains used concurrently in the interpretation process.

Popeye's domains

The bottom level domain (level (a) here) consists of program-generated "dotty" test pictures depicting a word made of laminar capital letters with straight strokes, drawn in a low resolution binary array with problems caused by overlap and artificially added positive and negative noise.

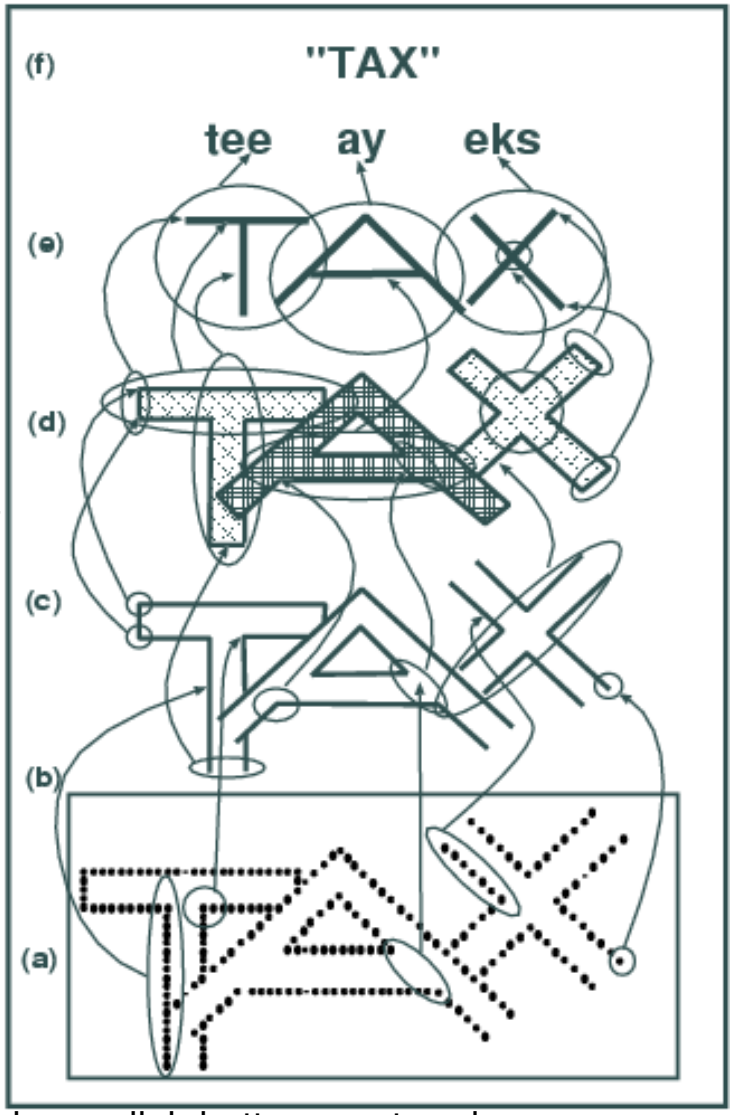
Collinear dots are grouped to provide evidence for lines in a domain of straight line segments, sometimes meeting at junctions and sometimes forming parallel pairs, as in (b).

The line, junction, and parallel-pair structures are interpreted as possibly representing (or having been derived from) structures in a 2.5D domain of overlapping flat plates (laminas) formed by joining rectangular laminar "bars" at junctions, as in (c).

The laminas are interpreted as depicting structures made of straight line segments meeting at various sorts of junctions, as found in capital roman letters with no curved parts – domain (d).

The components of (d) are interpreted as (possibly) representing abstract letters used as components of words of English, whose physical expression can use many different styles, or fonts, including morse code and even bit patterns in computers.

It degraded gracefully and often recognised a word before completing lower level processing.



The use of multiple levels of interpretation, with information flowing in parallel: bottom-up, top-down, middle-out, and sideways within a domain, allowed strongly supported fragments at any level to help disambiguate other fragments at the same level or at higher or lower levels.

(Note: A complete visual system would need many interfaces to action subsystems.)

Popeye's domains as dynamical systems

Although enormously simplified, and hand-coded instead of being a trainable system (this was done around 1975), Popeye illustrates some of the requirements for a multi-level perceptual system with dynamically generated and modified contents using ontologies referring to hypothesised entities of various sorts.

In the case of animal vision the ontologies could include

- retinal structures and patterns of many kinds
- processes involving changes in structures/patterns
- various aspects of 3-D surfaces visible in the environment: distance, orientation, curvature, illumination, shadows, colour, texture, type of material, etc.
- processes involving changes in any of the above e.g. changes in curvature if a something presses a soft body-part
- 3-D structures and relationships of objects including far more than visible surfaces, e.g. back surfaces, occluded surfaces, type of material, properties of material, constraints on motion, causal linkages.
- biological and non-biological functions of parts of animate and inanimate objects.
- actions of other agents.
- mental states of other agents,
- possibilities, and causal interactions increasingly remote from sensory contents of the perceiver,

Further development would require reference to 3-D structures, processes, hidden entities, affordances, other agents and their mental states.

All these ontologies would have instances of varying complexity, capable of being instantiated in dynamical systems whose components are built up over extended periods of learning, and which are mostly dormant unless awakened by perceptual input, or various types of thinking, e.g. planning the construction of a new toy.

Ideas used in Popeye were much influenced by conversations with Max Clowes.

Work to be done

The previous diagrams are not meant to provide anything remotely like an explanatory model: they merely point at some features that seem to be required in an explanatory model of the phenomena presented here and other aspects of animal vision (e.g. (Sloman, 2011)).

A more detailed account would have to explain

- exactly what is available in the dormant and active subsystem before each experiment starts;
- how learning can both extend old domains and produce new domains;
- what processes of awakening dormant sub-systems and propagating connections to revive other dormant systems occur;
- how those processes create a new temporary representation of the scene currently being perceived;

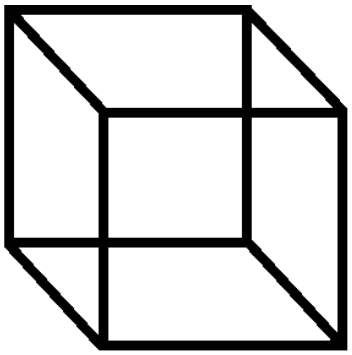
  (and in real life not just descriptions of a static scene, but of of ongoing processes – e.g. with people, vehicles, clouds, waves, animals, or whatever moving in various ways)
- what gets saved from previous constructions when a new scene is presented, and why that is saved, how it can be used, what interferes with saving, what removes saved items, etc.

I don't know of anything in machine vision, or detailed visual theory that comes close to this.

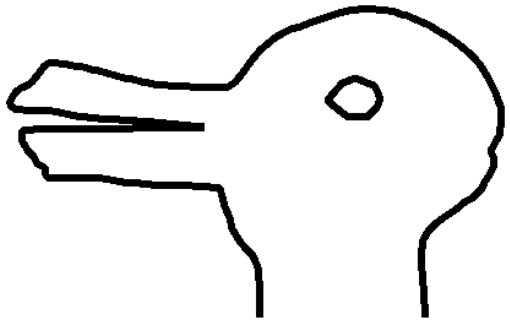
Perhaps the exceptions are the ideas of Julian Hochberg summarised in (Peterson, 2007) and Arnold Trehub's work in The Cognitive Brain, (Trehub, 1991) available online:
`http://www.people.umass.edu/trehub/`

More evidence for ontological variety

When ambiguous figures flip between different interpretations while you stare at them, what you see changing gives clues as to the ontologies used by your visual system.



Necker Cube



Duck-rabbit

When the left figure flips you can describe the differences in your experience using only geometrical concepts, e.g. edge, line, face, distance, orientation (sloping up, sloping down), angle, symmetry, etc.

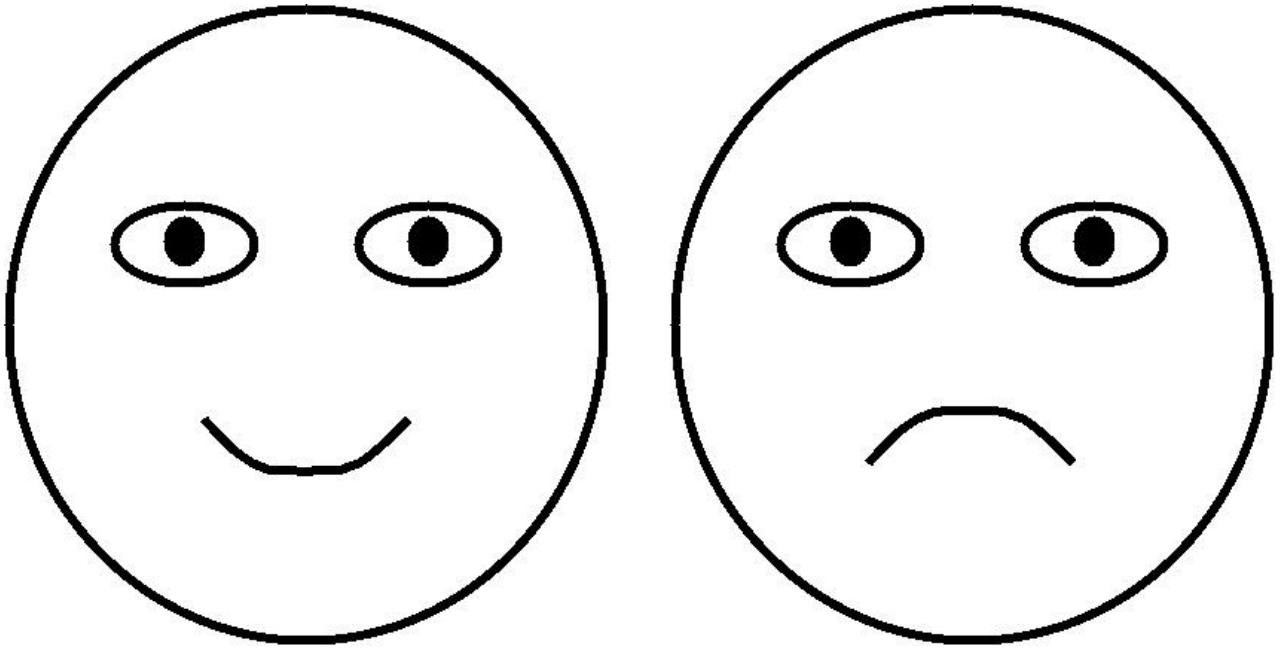
In contrast, when the figure on the right flips there is usually no geometrical change, and the concepts (types of information contents) required are concerned with types of animal, body parts of the animal and possibly also their functions. You may also see the animals as looking or facing in different directions.

Very different multi-stable subsystems are activated by the two pictures, but there is some overlap at the lowest levels, e.g. concerned with edge-features, curvature, closeness, etc.

Much richer ontologies are involved in some of the photographs presented earlier.

A more abstract ontology for seeing

Sometimes illusions (as opposed to ambiguous figures) give clues as to the ontology used in vision.



If the eyes in the right picture look different from the eyes in the left picture, then your visual system is probably expressing aspects of an ontology of mental states in registration with the optic array.[*]

The two pictures are geometrically identical except for the mouth.

---

[*] I have some ideas about why this mechanism evolved, but I leave it to readers to work out.

Process perception

What happens when the contents of perception are not static structures but processes in the 3-D environment, some, but not all, caused by the perceiver?

- A lot of intelligent perception is concerned with processes that are not occurring but in principle might occur.
- J.J. Gibson's theory of affordances is concerned with perception of possibility of certain processes that the perceiver could in principle initiate, even if they are not initiated.

  A more detailed presentation and critique of Gibson's view of functions of vision, compared with others is: Talk 93 What's vision for, and how does it work? From Marr (and earlier) to Gibson and Beyond.
  `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#gibson`
- But humans and many other animals can also perceive processes and possibilities for processes ("proto-affordances") that have nothing to do with their own actions or goals.

  E.g. seeing the possibility of a ball bouncing down a staircase, when the ball is on the top step.
- They can also see constraints on such processes: some of them only empirically discoverable, such as constraints on motion produced by wooden chairs and tables.
- Other constraints go beyond what is empirically detectable, but can be derived by reasoning, as in topological reasoning, or in Euclidean geometry.

  The mechanisms that allow minds (or brains) to transform knowledge acquired empirically into something more like a deductive system in which theorems can be proved, were probably the basis of the transition in language learning from pattern acquisition to a proper syntax.
  See `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#toddler`

Other Natural and Artificial Vision Challenges

The following present pictures with related challenges of different sorts:

- `http://www.cs.bham.ac.uk/research/projects/cogaff/challenge.pdf`
  Seeing various ways to re-stack cup saucer and spoon (Feb 2005)
- `http://www.cs.bham.ac.uk/research/projects/cosy/photos/crane/`
  Seeing and understanding pictures of a toy crane made from plastic meccano, and a few other things. (Jul 2007)
- `http://www.cs.bham.ac.uk/research/projects/cosy/photos/penrose-3d/`
  Variations on a theme: what to do about impossible objects. (Aug 2007)
- `http://www.cs.bham.ac.uk/research/projects/cogaff/challenge-penrose.pdf`
  More about impossible objects. (April 2007)

NOTE

When I first produced some of the above "challenge" presentations referring to impossible objects, I thought I was the first to use such pictures to argue that human vision does not involve constructing a consistent model of the scene.

However, Julian Hochberg got there first, around 40 years ago. See (Peterson, 2007)

and Mary Peterson's publications – some with Hochberg,
`http://www.u.arizona.edu/~mapeters/` (click on left of page)

Related presentations and papers

Additional related presentations are listed here

http://www.cs.bham.ac.uk/~axs/invited-talks.html

Related papers, old and new, can be found here

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

Including several papers on nature/nurture tradeoffs, by Chappell and Sloman.

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

Especially this:

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0801a

COSY-TR-0801 (PDF)

Architectural and representational requirements for seeing processes, proto-affordances and affordances.

Contribution to Proceedings of BBSRC-funded Computational Modelling Workshop,

Closing the gap between neurophysiology and behaviour: A computational modelling approach

Birmingham 2007, edited Dietmar Heinke

http://comp-psych.bham.ac.uk/workshop.htm

University of Birmingham, UK, May 31st-June 2nd 2007

Discussions of different functions of vision can be found in: (Sloman, 1986) (Sloman, 1989) (Sloman, 1994) (Sloman, 1996) (Sloman, 1998) (Sloman, 2002) (Sloman, 2001) (Sloman, 2005b) (Sloman, 2005a) (Sloman & CoSy project, 2006) (Sloman, 2006)

http://www.cs.bham.ac.uk/research/projects/cogaff/misc/triangle-theorem.html

Hidden depths of triangle qualia.

Different but related theories of vision

The idea that visual perception processes make use of different layers of interpretation is very old and takes many forms (including the idea of "Analysis by synthesis" in (Neisser, 1967)

Later work, e.g. by Irving Biederman proposed that 3-D visual perception could interpret objects in the environment as formed from combinations of object prototypes (e.g. "geons" in the case of (Hayworth & Biederman, 2006)).

Whereas the earlier systems (including Popeye) merely demonstrated the principle of multi-layer interpretation with all the mechanisms hand-coded, later systems learnt for themselves how to analyse images in terms of layered levels of structure. An example current "state of the art" system developed by Ales Leonardis and colleagues is summarised here: `http://www.vicos.si/Research/LearnedHierarchyOfParts`

In contrast the kind of vision system conjecture here uses layers of interpretation that are not based on part whole relationships, but on differences of domain, as illustrated in a relatively simple case by the Popeye program.

There is much more work to be done unravelling the multiple functions of vision, including controlling processes, understanding processes, understanding functional relationships between parts of complex mechanisms, understanding causal relationships, perceiving of various sorts of affordances, solving problems, and making plans.

Some hard to model visual capabilities involved in making discoveries in Euclidean geometry are discussed in:

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

More Pictures

More photographs by Jonathan Sloman are available here

`http://www.jonathans.me.uk/index.cgi?section=picarchive`

References (to be extended)

References

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

Frisby, J. P. (1979). *Seeing: Illusion, brain and mind*. Oxford: Oxford University Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Hayworth, K. J., & Biederman, I. (2006). Neural evidence for intermediate representations in object recognition. *Vision Research (In Press)*, *46*(23), 4024–4031. (http://geon.usc.edu/~biederman/publications/Hayworth_Biederman_2006.pdf)

Marr, D. (1982). *Vision*. San Francisco: W.H.Freeman.

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.

Peterson, M. A. (2007). The Piecemeal, Constructive, and Schematic Nature of Perception. In M. A. Peterson, B. Gillam, & H. A. Sedgwick (Eds.), *Mental Structure in Visual Perception: Julian Hochberg's Contributions to Our Understanding of the Perception of Pictures, Film, and the World* (pp. 419–428). New York: OUP.

Sloman, A. (1971). Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd ijcai* (pp. 209–226). London: William Kaufmann. Available from http://www.cs.bham.ac.uk/research/cogaff/04.html#200407

Sloman, A. (1978). *The computer revolution in philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press). Available from http://www.cs.bham.ac.uk/research/cogaff/crp

Sloman, A. (1986). *What Are The Purposes Of Vision?* Available from http://www.cs.bham.ac.uk/research/projects/cogaff/12.html#1207 (Presented at: Fyssen Foundation Vision Workshop Versailles France, March 1986, Organiser: M. Imbert)

Sloman, A. (1989). On designing a visual system (towards a gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, *1*(4), 289–337. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#7

Sloman, A. (1994). How to design a visual system – Gibson remembered. In D.Vernon (Ed.), *Computer vision: Craft, engineering and science* (pp. 80–99). Berlin: Springer Verlag. Available from info:vKA5pOB9aq8J:scholar.google.com

Sloman, A. (1996). Actual possibilities. In L. Aiello & S. Shapiro (Eds.), *Principles of knowledge representation and reasoning: Proc. 5th int. conf. (KR '96)* (pp. 627–638). Boston, MA: Morgan Kaufmann Publishers. Available from http://www.cs.bham.ac.uk/research/cogaff/96-99.html#15

Sloman, A. (1998, August). Diagrams in the mind. In *in proceedings twd98 (thinking with diagrams: Is there a science of diagrams?)* (pp. 1–9). Aberystwyth.

Sloman, A. (2001). Evolvable biologically plausible visual architectures. In T. Cootes & C. Taylor (Eds.), *Proceedings of British Machine Vision Conference* (pp. 313–322). Manchester: BMVA.

Sloman, A. (2002). Diagrams in the mind. In M. Anderson, B. Meyer, & P. Olivier (Eds.), *Diagrammatic representation and reasoning* (pp. 7–28). Berlin: Springer-Verlag. Available from http://www.cs.bham.ac.uk/research/projects/cogaff/00-02.html#58

Sloman, A. (2005a, September). *Discussion note on the polyflap domain (to be explored by an 'altricial' robot)* (Research Note No. COSY-DP-0504). Birmingham, UK: School of Computer Science, University of Birmingham. Available from http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0504

Sloman, A. (2005b). *Perception of structure: Anyone Interested?* (Research Note No. COSY-PR-0507). Birmingham, UK. Available from http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0507

Sloman, A. (2006). Polyflaps as a domain for perceiving, acting and learning in a 3-D world. In *Position Papers for 2006 AAAI Fellows Symposium.* Menlo Park, CA: AAAI. (http://www.aaai.org/Fellows/fellows.php and http://www.aaai.org/Fellows/Papers/Fellows16.pdf)

Sloman, A. (2011, Sep). *What's vision for, and how does it work? From Marr (and earlier) to Gibson and Beyond.* Available from http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk93 (Online tutorial presentation, also at http://www.slideshare.net/asloman/)

Sloman, A., & CoSy project members of the. (2006, April). *Aiming for More Realistic Vision Systems* (Research Note: Comments and criticisms welcome No. COSY-TR-0603). School of Computer Science, University of Birmingham. (http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0603)

Trehub, A. (1991). *The Cognitive Brain*. Cambridge, MA: MIT Press. Available from http://www.people.umass.edu/trehub/

(Etc.)