

# INCOMPLETE DRAFT

(Last revised November 29, 2005)

## What the brain's mind tells the mind's eye.

Aaron Sloman

School of Computer Science, University of Birmingham

Birmingham, B15 2TT, UK

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

November 29, 2005

### Abstract

**[Abstract is now out of date. Needs to be rewritten. The new version of this paper includes a new (partial?) taxonomy of types of function of vision, showing that different forms of representation and different mechanisms are involved.]**

Clearly we can solve problems by thinking about them. Sometimes we have the impression that in doing so we use words, at other times diagrams or images. Often it feels as if we use both. Sometimes we have no idea what we are doing. What is going on when we use mental diagrams or images? This question is addressed in relation to the more general multi-pronged question: what are representations, what are they for, how many different types are there, in how many different ways can they be used, and what difference does it make whether they are in the mind or on paper? The question is related to deep problems about how vision and spatial manipulation work. We are far from understanding what is going on. In consequence of our not understanding this we cannot design user interfaces that understand their displays in the same way as human users do. In particular we need to explain how people (and some other animals) understand spatial structure and motion, and how we can think about objects in terms of a basic topological structure with more or less additional metrical information. I shall try to explain why modelling human visual perception is a problem with hidden depths, since we do not really know what all the functions of vision are, and in some cases even when we have labels (e.g. Gibson's notion of "perception of affordances" ) it can be hard to explain what they mean. I shall suggest that our grasp of spatial structure and affordances inherently includes a grasp of a complex range of possibilities and their implications (counterfactual conditionals), many of them indexed by locations in a scene or parts of seen objects, and also by possible goals. Different sorts of examples need to be analysed in great depth in order to identify requirements for human visualisation capabilities. Some involve 2-D structures and processes, some 3-D. Some involve continuous structures and processes, others discrete ones, and some both. Some involve finite structures or sequences, others infinite ones. It is all far beyond the current state of the art in machine vision. However, it should be possible to make progress by analysing the problems carefully and choosing designs accordingly.

[CHANGES: 29 Nov 2005: I have not made any changes to this paper for over a year. However, I have been thinking and working on the problems, and, partly as a result of working on the CoSy project (referred to in Section 2.2) I now see vision (and more generally perception) as primarily concerned with perceiving and understanding *processes* as opposed to *structures*. Some of the implications of this ‘gestalt-switch’ are presented in two PDF slide presentations which partly overlap with this paper:

<http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0505>

A (Possibly) New Theory of Vision

<http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0506>

Two views of child as scientist: Humean and Kantian

That work arose in the context of work on requirements for representation in the CoSy robot scenarios, available here

<http://www.cognitivesystems.org/files/dr-02-01-rev1.pdf>

<http://www.cs.bham.ac.uk/research/projects/cosy/papers/index.php#tr0507>

21 Oct 2004: Slightly expanded ‘Continuous *vs* enduring’ section 3.4, with comment on discrete sampling of continuous motion or surfaces. Moved point about virtual machine functionalism to footnote.

27 Sep 2004: added a couple of paragraphs at end of ‘prologue’. (Section 1). Added section ‘Continuous *vs* enduring’ section 3.4

26 Sep 2004: reduced top margin to avoid problem of loss of text at bottom in PDF version reported by John Knapman.

19 Sep 2004: Add comment about identifying functions between logically prior to identifying mechanisms, in 3. Expanded 3.8.2 on ‘episodic’ memory. Included more on the differences between continuous and discrete information structures, showing how this is orthogonal to the distinction between implicit and explicit information.

12 Sep 2004: added the section on pre-cursors of episodic memory (section 3.8.2) and the distinction between ‘perceiver-centred’ and ‘vicarious’ perception of affordances (section 3.15).

]

## Contents

<b>1</b>	<b>Prologue</b>	<b>5</b>
<b>2</b>	<b>Why is Human-Machine Symmetry so far off?</b>	<b>5</b>
2.1	Is natural language understanding easier? . . . . .	6
2.2	Obstacles to visual understanding . . . . .	7
<b>3</b>	<b>Varieties of vision</b>	<b>8</b>
3.1	Explicit and implicit information – and other distinctions . . . . .	9
3.2	Structure <i>vs</i> Affordances . . . . .	10
3.3	Discrete <i>vs</i> continuous <i>vs</i> fuzzy . . . . .	10
3.4	Continuous <i>vs</i> enduring . . . . .	11
3.5	Implicit prediction for control . . . . .	12
3.6	Multi-layered vision . . . . .	13
3.7	Varieties of affordance . . . . .	13
3.8	Vision in action: transient, implicit vision . . . . .	14

3.8.1	Perceiving a static scene while eyes and head move . . . . .	16
3.8.2	Precursors of ‘episodic memory’ . . . . .	16
3.8.3	Implicit information and dynamical systems . . . . .	17
3.9	Varieties of learning . . . . .	18
3.10	Explicit associative perceptual information . . . . .	19
3.11	Non-transient direct perceptual information . . . . .	20
3.11.1	Perceiving structures and perceiving affordances . . . . .	20
3.12	Perceiving perception: self-awareness . . . . .	22
3.13	Attention . . . . .	22
3.14	Aesthetic aspects of perception . . . . .	23
3.15	Perceiver-centred <i>vs</i> vicarious perception . . . . .	23
3.16	Summary so far . . . . .	24
<b>4</b>	<b>Perceiving structures, at different levels of abstraction</b>	<b>24</b>
4.1	Larger 2-D structures and more abstract 2-D structures . . . . .	26
4.1.1	Agglomerations in 2-D fields . . . . .	26
4.1.2	Abstract features perceived in 2-D fields . . . . .	27
4.2	Geometric and non-geometric 3-D spatial interpretations . . . . .	28
4.3	Combining structural and procedural information . . . . .	30
4.4	Proto-deliberative reactive perceivers . . . . .	31
4.5	Deliberative perceivers . . . . .	31
4.6	Storing information for rapid access . . . . .	32
<b>5</b>	<b>What is seen: clues from different sorts of ambiguity</b>	<b>33</b>
5.1	Perceiving non-geometrical features, states, processes . . . . .	33
5.2	Seeing or inferring? . . . . .	34
<b>6</b>	<b>Visions of vision</b>	<b>36</b>
6.1	Seeing according to Marr . . . . .	37
6.2	What Marr left out: Gibsonian affordances and mental states of others . . . . .	37
6.3	Perceiving empty spaces . . . . .	39
<b>7</b>	<b>The “contents” of visual perception</b>	<b>39</b>
<b>8</b>	<b>Betty Crow: cognitive agent and hook-maker</b>	<b>42</b>
8.1	What sort of architecture could do what Betty did? . . . . .	42
8.2	What should amaze us? . . . . .	43
8.3	What a child cannot see . . . . .	44
8.4	Vision and affordances . . . . .	45
<b>9</b>	<b>Humans can think with spatial structures.</b>	<b>46</b>
9.1	An example . . . . .	46
9.2	Can we believe introspections? . . . . .	46
<b>10</b>	<b>How can we increase HMS?</b>	<b>48</b>

**11 Simple examples** **49**  
11.1 Varieties of limitations of current machine image understanding . . . . . 49  
11.2 What is it to understand the scene? . . . . . 49

**Acknowledgements**

**References**

# 1 Prologue

This paper was started in response to an invitation in 2002 to contribute to a new journal on diagrams, originally due out in 2003. It was going to be a paper on how our ability to see and use diagrams was related to our ability to perceive affordances. As I worked on the paper I gradually became aware of more and more gaps and deficiencies in my analysis (including work on vision previously published in chapter 9 of Sloman (1978)<sup>1</sup> Sloman (1989), and Sloman (2001b), and work on representation previously published in Sloman (1971) and several sequels to that paper).

The need for a deeper richer theory, combined with the pressure of other work, caused repeated delays in completion. Even now it is not complete. Perhaps it never will be.

This draft is made available on the web as an invitation to others to contribute to the conceptual clarification and design work that needs to be done. One of the main changes in my thinking about the paper came from a realisation that many of the points could be made clearly and effectively only in the context of a fairly comprehensive taxonomy of *functions* of vision, not only in humans or human-like robots, but in all varieties of animals and machines. In part that is because most of those functions, including evolutionarily very old functions shared by far less sophisticated animals, remain among the diverse functions of human vision. Thus even the most ‘sophisticated’ seers still use very primitive visual mechanisms operating in parallel with and sharing sub-mechanisms (e.g. retinal cells) with advanced human visual processes, such as reading text or musical scores, understanding diagrams, seeing how a machine works and seeing someone looking disappointed. Moreover, these differences in visual function are also related to differences in ways in which information is encoded and used, i.e. to differences in forms of representation.

Some of the differences in function are so great that some researchers prefer to drop the word ‘representation’ for some cases: but that just leads to silly semantic debates about what the word ‘representation’ means, diverting attention from the deep scientific and engineering issues about the variety of ways in which information can be stored, transmitted, analysed, transformed, searched, interpreted and used. That’s like arguing over whether an isotope of carbon should be called ‘carbon’ instead of getting on with understanding the periodic table of the elements.

So instead of arguing about what is and is not a representation we should collaborate in the analysis of varieties of types of representation, and their functions. When we understand that variety properly we shall see that some debates, e.g. between those who favour symbolic and those who favour sub-symbolic mechanisms miss the point that both are needed in some organisms and robots. Likewise debates about the role of embodiment and about dynamical systems vs computational systems are pointless because both sides ignore some of the variety of phenomena that we need to understand.

Perhaps unification of a diverse research community is too much to hope for, but it could be a side-effect of what follows.

## 2 Why is Human-Machine Symmetry so far off?

Currently there is glaring asymmetry between computers and humans even when they interact using some structure that the computer in some sense ‘understands’, e.g. a form or a mouse-driven

---

<sup>1</sup>Now available online at <http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html>

interface, or certain classes of graphical tools.

For example, when computers display forms for a human to fill in there may be instructions or questions on the form which the human understands but the computer does not, and the human can provide answers that the computer does not understand, though it may be able to manipulate them and store them in a form that can be processed by the computer and then present new derived information in a form that is understood by a human but not the computer. The computer may even be able to perform statistical analyses of large numbers of responses, yet have no idea what statistical methods do, or what the results mean, even when they have been transformed, by the computer itself, into a graphical representation that humans find very useful.

More generally, computers may be able to do syntactic manipulations of verbal and other structures, where the manipulations are meaningful to us though the computer has no understanding of their import because it lacks a semantic interpretation of the structures.

I am not following Searle (1980), and many others, in claiming that machines, or computer-based machines, can *never* understand things as humans do, a claim that is often driven more by wishful thinking than by argument.<sup>2</sup> I am merely stating the obvious: so far non-biological machines do not understand sentences and pictures as humans do, though it is important to notice that even if they do not interpret the structures they are manipulating as we do they must (as pointed out in Sloman (1985)) have *some* understanding of those structures in order to be able to manipulate, transform, compare, combine, store, and search for them. However, by analysing in depth what is missing we may be able to move closer to adequate designs for machines that share *our* understanding. This will also help us understand what the capabilities of humans are, and how they can be explained, continuing a project outlined in Sloman (1985, 1987).

## 2.1 Is natural language understanding easier?

To some extent, advances in AI research on natural language understanding may reduce the lack of symmetry when the interface is entirely linguistic, though there are still formidable problems, since understanding linguistic communications generally requires deep domain knowledge (as pointed out half a century ago by Bar-Hillel). A child cannot understand ‘John went to Mary’s party’ without using a great deal of knowledge about space, time, motion, what human beings are, what parties are, what they are for, who John is and who Mary is. Both may be either real individuals or story characters. Understanding the last case includes understanding why the real John who lives next door cannot go to the party in the story. These are all hard problems for AI – and psychology.

Nevertheless there are many people working on use of natural language by machines, and it is likely that useful natural language interfaces can be developed relating to limited domains where the computer has quite a lot of information about that domain, e.g. diagnosis and treatment of a particular class of diseases or fault diagnosis in a fairly simple machine, just as we can produce machines that play chess at least as well as most humans. It may be possible for the computer to get by in restricted contexts with partial understanding of what it is talking about, like a human clerk who can help someone by looking up and reporting or changing the contents of books, diagrams, tables, etc. without really understanding much about the subject matter. In such cases it is not hard to probe a little and find that the understanding between human and machine, or between employer

---

<sup>2</sup>Bar-Hillel’s famous paper on the impossibility of fully automatic machine translation makes a more focused claim (Bar-Hillel, 1964).

and clerk is not symmetric. Asking the clerk, or the machine, a slightly unusual question ('Did any other disease respond to our proposed treatment?') may produce incomprehension even though the answer is readily accessible in the pages recently examined. Many will argue that, despite sharing such limitations with a machine, the human assistant, unlike any machine we currently know how to design, has a full understanding of many aspects of the communication, going far beyond 'mechanical' following of a complex collection of rules.

It is sometimes thought that by embedding linguistic mechanisms in the context of a 'complete' machine with the ability to perceive and act on the world we can replicate such understanding (the frequently re-invented 'systems' reply discussed and criticised by Searle in his 1980 paper). However, there are indefinitely many ways of producing any behaviourally specified competence, no matter how many varieties of input and output the system has, and many of them will not provide human-like understanding no matter how convincing and human-like their externally visible behaviour. *How* the behaviour is produced is also important – and that depends on the information-processing architecture.<sup>3</sup>

This paper explores some of the requirements for such an architecture in the context of visual understanding, but the implications are more general. For example, we shall see that using vision to perceive affordances involves the ability to use information about what does not exist but might exist. This is related to the ability to understand descriptions of what exists but is not perceived, statements about the past, statements about the future, and fictional stories that are not about the past, the unseen present, or the future (Dennett, 1987). However, we shall see that there are simpler forms of understanding, shared with many other organisms, which do not require such architectures, some of which use only implicit representations.

## 2.2 Obstacles to visual understanding

Many deep difficulties stand in the way of achieving symmetric human-machine communication based on diagrams, images, or a shared visual viewpoint in space. A similar lack of symmetry can occur between an adult and a child looking at the same physical objects, as illustrated below in Section 8.3

In order both to explain some of the limits common to current AI vision systems and psychological theories about the nature of vision, and to suggest directions for future work, this paper presents an analysis of types of things that can be seen by different sorts of organisms. This amounts to an (incomplete, first draft) analysis of functional requirements for human vision, which overlaps with requirements for many other organisms, and future robots. The analysis is guided by design considerations, referring not just to differences in content of what is perceived, or differences in physical environments, but also differences between the forms of representation required and the mechanisms required for different visual tasks or sub-tasks.

In part the aim is to provide a framework which shows that some disputes concerning how to model or explain visual capabilities arise out of the fact that different researchers focus on mechanisms required for different sub-functions of vision. So some disagreements, for instance disagreements as to whether vision should use distributed representations in neural nets or hierarchically structured symbolic descriptions, are pseudo-disagreements, since the proponents are arguing at cross-purposes. They are all right about what may be needed for some visual

---

<sup>3</sup>See Sloman (1992)

subtasks, and all wrong about what suffices for all visual processing. This is closely related to the discussion of ontological blindness in Sloman and Chrisley (2004), which describes how scientists and engineers may be ‘blind’ to some aspects of the things they study through use of inadequate ontologies.

Unfortunately, our understanding of the nature of normal adult human vision is still very limited and as a result we do not even know what the requirements are for a machine that shares our visual capabilities. I have previously tried to identify some of the problems in Sloman (1978)[chapter 9], Sloman (1989) and Sloman (2001b) but the problems are deeper than those papers indicate. I shall try to explain why that is so by analysing a variety of requirements for human-like visual perception (some but not all shared with other animals) along with some tentative partial proposals for mechanisms that can fulfil those requirements. I shall attempt to draw out features of vision that are ignored or unnoticed by most vision researchers, especially vision that involves acquisition of explicit, reusable information about certain subtle aspects of the environment often referred to as ‘affordances’, following Gibson, though I have no idea whether Gibson ever thought about some of the types of affordance presented below, or whether he would accept what is said about them here.

In order to explain what this means I shall contrast seeing affordances with other types of vision requiring different mechanisms and different forms of representation. In addition, *seeing* affordances can be contrasted with *inferring* affordances from what is seen. That distinction can be contrasted with a distinction between *implicit* and *explicit* perception of affordances, where the latter requires mechanisms and forms of representation not required for the former.

Section 3 presents a taxonomy of types of functions of vision distinguished according to the forms of representation required, as summarised in Figure 1. Later sections will elaborate on a subset of the categories, especially the explicit perception of affordances, and will discuss additional categories. No claim is made about completeness. I expect we shall still be learning surprising things about vision for decades to come. Some of the gaps will be revealed by attempting to design working robots that combine all the varieties of function described here.<sup>4</sup>

### 3 Varieties of vision

Psychologists and neuroscientists have begun to identify different sorts of visual pathways in brains, e.g. Goodale and Milner (1992). There is, however, a logically prior task of identifying different *functions* of vision on the basis of which we can interpret evidence about neural mechanisms.<sup>5</sup> A rather shallow attempt to do that is the well known suggestion that some visual pathways are concerned with ‘what’ (i.e. object identification) others with ‘where’ (i.e. object location). In this

---

<sup>4</sup>A subset will be attempted in the multi-site EU-funded CoSy project, due to start in September 2004, described here <http://www.cs.bham.ac.uk/research/projects/cosy/>.

<sup>5</sup>It is *logically* prior in the sense that unless the current ontology used by researchers includes a visual function **VF** they will not be able to ask whether a particular neural mechanism does or does not contribute to function **VF** and they may even ignore some mechanisms because they clearly do not serve any of the currently identified functions. However, it need not be *temporally* prior: investigation of physiological mechanisms can lead to indications that some visual function has not been identified yet, or that what was previously regarded as a unitary function may actually involve two separate functions. For instance, seeing the spatial relation between a slit and a card held in the hand may appear to be one task. However, either task analysis or investigation of brain mechanisms can reveal that seeing the relation as part of the process of a feedback loop guiding insertion of the card into the slot is a quite different function from being able to inform someone else how the orientations differ, or being able to plan in advance the rotation required to make the card fit into the slot – as explained below.

paper I attempt to dig deeper into the variety of functions to be expected in sophisticated visual systems, especially human-like visual systems, building on earlier work in Sloman (1989), inspired by Gibson (1986)). This section presents a high level overview of some of the variety, summarised below in Figure 1. Later sections present more detailed analysis of some of the more complex functions, for instance those that involve perception of structure or perception of affordances, or perception of multiple affordances attached to various parts of complex structures.

### 3.1 Explicit and implicit information – and other distinctions

We start by distinguishing transient visual information used only *implicitly* in controlling actions from more enduring information *explicitly* represented in a form that can, in principle, be used for more than one purpose and can, in principle, survive both the sensory stimuli producing the information and the immediately triggered responses. This will be used as a basis for distinguishing implicit and explicit visual perception of affordances. Debates about which type is needed by an intelligent (or human-like) system are pointless if both are needed, for different reasons. We also introduce a variety of further sub-divisions corresponding to the acquisition of different kinds of information, or different ways of using visual information. Not all organisms, or robots, will have all varieties.

Processes producing the explicit visual information can be further subdivided into (a) *associative* or *indirect* perception, where the information gained depends on inferences using previously learned generalisations linking different phenomena and (b) more *direct* perception, e.g., where the explicit information acquired in vision is directly caused by and structurally related to what the information refers to. For instance, seeing a tree swaying in strong winds uses direct perception whereas seeing that there has been a storm by seeing its after-effects, or by reading a report about it, uses indirect, associative perception.

Within associative perception we can distinguish cases where the inferences make use of *general-purpose* reasoning mechanisms that can combine stored generalizations with new sensory information to derive consequences, from cases where *special-purpose* visual mechanisms, produced by design, evolution, or training, react very rapidly to particular sorts of perceptual inputs by generating interpretations.

The former (cases of associative perception using general-purpose reasoning mechanisms) are not regarded as visual processes (even though we often describe them by saying things like ‘I now see that the window must have been broken’).

The latter (cases of more direct associations detected by mechanisms specific to analysing and interpreting the contents of the optic array) can be described as ‘visual’ because they are dedicated to the analysis and interpretation of visual input and may form intermediate stages in some of the more complex uses of vision.

Associative perception is in some ways analogous to implicit perception, except that in the implicit case perceptual stimuli trigger particular actions (which may be external or internal), whereas in the case of special-purpose associative perception what is triggered is creation of some explicit information structure that can be used in different ways for different subsequent purposes, e.g. seeing different larger wholes containing the detected feature. E.g. a vertical line in the scene may be seen as an occluding edge of an object or as the edge at which two visible faces meet, or as a crack in a surface, depending on the rest of the visual input; and something can be recognised as

an eye or a tooth, and then subsequently seen as part of a tiger, or some other animal, depending on the visual context. (Experts may be able to discriminate the whole animal on the basis of smaller parts than non-experts.)

Like the implicit perception in reactive mechanisms, the special-purpose associative mechanisms may be either innate (or pre-designed in machines) or produced by training (e.g. skills based on ‘over-learning’). Many athletic skills use highly-trained implicit perception, whereas learning to read, or understand speech, fluently involves explicit perception, because whatever is read or understood can be used for different subsequent purposes by different sub-systems.

### 3.2 Structure vs Affordances

Within the sub-category of explicit direct (non-associative) perception we shall distinguish *perception of structure* and *perception of affordances*. The former is concerned with what *exists*, the latter with what is and is not *possible*, i.e. which actions could or could not exist in the perceived situation. (Compare the discussion of ‘actual possibilities’ in Sloman (1996a).) However, perception of affordances often intimately combines associative and direct perception, as will be explained below.

As argued in Sloman (1989) the low-level feature detectors that record various kinds of edge-features, optical flow patterns, texture, colours, specularity, etc. can be seen as examples of special-purpose associative mechanisms, since they *infer* new kinds of information from sensory information, though they do not use general-purpose inference methods and they do not use explicit generalisations as premises: instead the generalisations are ‘compiled into’ the mechanisms, often by evolution, though they may be trainable or tunable in adaptive sensor systems.

### 3.3 Discrete vs continuous vs fuzzy

Among the other distinctions that will be seen to be important is the distinction between ‘discretized’ and non-discretized information. For example, a visual feedback control system that continuously registers the relation between an object and the hand that is moving to grasp the object will have (to a first approximation, ignoring discrete detection of photons) a *continuously* varying information state. In contrast when a visual system categorises something as either (a) within reach, (b) reachable if the body moves, or (c) totally out of reach, it is using *discretized* information because information is ‘chunked’ into distinct categories that may be the basis of quite different actions. Discretization is essential for learning associations that can be used in multi-step planning and reasoning processes, as explained in Sloman et al. (2004).

As Zadeh and others have pointed out (Zadeh, 2001), discrete categories need not be sharply distinguished – he refers to ‘fuzzy granularity’. Chunked information structures, whether fuzzy or not, lend themselves to recursive composition of structures to express more and more complex contents. In contrast continuous information values are, at present, more often treated as components of fixed dimensional arrays or vectors – a form of information composition with which many mathematicians and engineers are comfortable.

Whether implicit information structures are continuous or discrete will depend on whether the dynamical systems in which they occur change only continuously or have separate attractors between which the system can switch. Explicit information structures may also be continuously

variable, for instance measures of degree of hunger or thirst that are accessible by several mechanisms and used for different purposes (Scheutz and Schermerhorn, 2002). Most symbolic AI systems making use of explicit information structures use discrete structures, but there is no reason why they should not also use continuously varying information structures, for instance if a symbolic rule uses the value of some continuously varying state to check a condition.

It is sometimes thought that a crucial difference between linguistic and pictorial (or analogical) forms of representation is that the former are discrete and the latter continuous. This is clearly not the case, as pointed out in Sloman (1971), since for instance a discrete sequence of names or descriptions can analogically represent an ordered set of objects or events, and moreover a discrete structure such as a circuit diagram, where only the topology, not the metrical relationships are important, is typically an analogical representation of a circuit.

### 3.4 Continuous vs enduring

Proponents of embodied cognition, or of the dynamical systems approach to mind

**[REFS: e.g. Special issue on Situated and Embodied Cognition, Editor Tom Ziemke, Cognitive Systems Research, 3,3 Dec 2002, <http://www.elsevier.com/locate/cogsys> ]**

point out that many human activities depend on the fact that behaviour is not simply controlled by central decision-making brain mechanisms, but often involves closed causal loops in which continuously changing states of the body, objects in the environment, sensors, central brain mechanisms, and motor subsystems interact. A child standing on a swing and making it gradually swing faster is using a host of subtle feedback loops both to control balance and to control the timing, direction and magnitude of forces exerted on the seat, on the child's torso (e.g. when knees are straightened) on the ropes, etc. Anyone who has learnt to play the violin finds that getting a good tone while playing a single soft note requires months or years of teacher-guided practice to coordinate sensors in fingers, arm and ear, control of pressure on the bow, the direction of motion of the bow, the angle at which the bow is held, the detailed movements of upper and lower arm, etc. Similar kinds of tight integration between brain, body and environment are involved in the diving and swooping of insect-catching birds, the rapid movements of a squirrel or monkey along and between flexible branches in trees, and many kinds of highly skilled operation of complex machines from scythes to jet fighters.

All of that is true and important and certainly needs to be understood by researchers who hope to explain how such animals work, or to produce robot swing-users, violin-players, swallows, squirrels, monkeys, and fighter pilots.

**[REFS: <http://news.bbc.co.uk/1/hi/sci/tech/628270.stm> (Robo Monkey)]**

However, there are three points to be made about this:

1. Not all of human intelligence is of this sort (e.g. thinking about why the number of primes less than  $X$  divided by  $X$  will decrease as  $X$  increases, or thinking about how many different twelve-tone musical sequences there are, or composing a story in your head).
2. The fact that a dynamical system includes (to a first approximation in a quantum physical universe) continuously changing physical components in the environment does not require everything in the central control system to change continuously (e.g. witness the trade-offs between digital and analog designs for radio receivers).
3. Sometimes what is important for dealing with a continuously changing environment is not that something in the brain (or artificial control system) changes continuously, but that there

is an *enduring* representation of the continuously changing entity with the right causal powers within the system. For instance, the enduring representation may change at intervals because states of the continuously changing object are *sampled* at intervals, just as discrete optical sensors can sample an image at spatial intervals even if the image varies continuously. The processes of interpretation of such discrete image data can use the fact that the discrete image representation (e.g. some sort of array of pixel values) is a representation of something spatially continuous, for instance by hypothesising lines or edges in the scene whose locations are specified with sub-pixel accuracy. Similarly even though the intervening states of a continuous process are not continuously sensed, some of the non-sensed intermediate states may be inferred when appropriate, on the basis of records of discrete changes in the enduring representation. So when you watch a lion chasing a deer there are two independently moving objects between which your attention may be switching (including saccadic shifts) but despite discrete sampling you interpret the environment as containing two continuously, independently, moving objects.

Point (3) is particularly important here and will be followed up in section 3.8.2 on precursors of episodic memory below. In particular, an organism or robot that deals with things in the environment that endure between moments when they are sensed may need to retain information about those objects during periods when they are not sensed. That information can be used for different purposes, e.g. predicting where they will be sensed later, planning capture or avoidance actions, trying to form generalisations about how the external entity behaves, etc. Such enduring representations need not be under the continuous control of sensors, nor do they have to exist in a faithful continuously changing internal model of the environment, for there are many more economical ways of doing prediction and planning.<sup>6</sup>

### 3.5 Implicit prediction for control

Prediction can occur not only within a deliberative mechanism exploring possible futures, but also within the low level workings of a reactive control system. For example, during bipedal running various signals to motors cause a foot in contact with the ground to push the ground in a manner that launches the body forward and upward. The other foot needs to be in position to absorb the shock of landing and start pushing in the next stride. This requires a subtle and rapid change in muscles from being slightly relaxed (absorbing shock) to being very tense starting the next stride. If the time to contact can be predicted the sequence of changes when contact is made can be prepared in advance. This does not necessarily mean that there is anything like an explicit description of the future event: ‘The foot will make contact in 200 milliseconds’ or whatever. Rather the prediction can be implicit in a variety of processes including the speed at which the the foot not in contact with the ground is brought forward and the distance it is brought forward. That prediction could either be a result of individual learning which adjusts control parameters, or a result of evolution (species learning) which produces a physical configuration of bone, muscle, skin, etc., and pre-wired control patterns. Either way it involves implicit information about what is going to happen and when. There are many other examples where smooth, fast, skilful action requires prediction, which can be either implicit

---

<sup>6</sup>This use of multiple enduring, but independently changing, representations is connected with the differences between ‘Atomic state functionalism’, which regards mental processes as sequence of indivisible states, like a finite-state automaton, and ‘Virtual machine functionalism’ which regards mental processes as involving arbitrarily many concurrently active, enduring, interacting components (Sloman, 1993).

### 3.6 Multi-layered vision

Perception often requires multiple stages where features, objects or relations detected in some intermediate stage form cues for subsequent stages, or are interpreted as parts of some larger structure. In some cases the links between individual cues and information items triggered by the cues may be ‘associative’ in the sense defined above, i.e. based on an arbitrary generalisation (i.e. not something derivable by reversing the image formation process), whilst the *organisation* of the information thus built up is ‘direct’ because the structures are built in registration with the sensory input, e.g. different parts of the derived information structures may be mapped onto locations in the visual field, or onto locations in a general-purpose representation of the external spatial environment. Thus seeing a person as ‘exerting effort’ involves a very abstract non-spatial description, but the effort may be perceived as located in a part of the scene, where the person is pushing, or pulling something, for instance.

There is no requirement for such multi-layered processing to use a uni-directional pipeline, with all information flowing ‘upwards’ from sensors to more complex and abstract information structures. On the contrary, there is much evidence of ‘top-down’ and also ‘sideways’ (including cross-modal) influences in vision, auditory perception, haptic perception, etc., and one reason for this is that such co-operative processing can allow global information to resolve local ambiguities more quickly, by adding constraints on interpretations consistent with all available information.

Neither should we assume that all the intermediate perceptual structures form a linear structure (e.g. a bi-directional pipeline). On the contrary there are good reasons for believing that there are many branching directions of influence in perceptual systems, e.g. because different perceptual tasks may share some sub-tasks. This was described as ‘labyrinthine’ perception in Sloman (1989). The key ideas are much older, and can be found in the idea Neisser proposed in the 1960s of ‘analysis by synthesis’, sometimes referred to as ‘hierarchical synthesis’ (Neisser, 1967).

These are all somewhat crude distinctions with many intermediate or hybrid cases, some of which will be mentioned below.

### 3.7 Varieties of affordance

The conceptual framework sketched in the last few paragraphs will be expanded below, and summarised in Figure 1. We shall see in later sections that there are two importantly different ways in which an animal or machine may use information acquired perceptually, namely by *implicitly* encoding the information in the transient patterns of activity produced by sensory stimulation, or by *explicitly* recording information independently of how it is used. This leads to two notions of ‘affordance’, namely the affordances involved in transient implicit vision and the affordances involved in non-transient, explicit vision. We may call them ‘implicit affordances’ and ‘explicit affordances’ respectively. The implicit affordances are closely related to what are often referred to as ‘sensori-motor contingencies’ (O’Regan and Noë, 2001), and are more basic than explicit affordances in the sense that they are relevant to a wider variety of types of animals and are

evolutionarily older.

In purely reactive organisms where perception immediately causes (internal or external) actions without any intervening plan-construction and without formation of explicit intentions to perform actions prior to performing them, it could be said that *all* perception is implicit perception of affordances: there is no other kind of perception. That is not true for all organisms.

This paper is more concerned with explicit affordances, as a special case of perception that provides explicit information that is separable from the processes that use the information. However we also need to understand implicit affordances because those are the only ones considered by some researchers who deny the need for explicit representations in intelligent systems. From our point of view, both play a role and it is silly to claim that only one sort ever exists.

The rejection of symbolic AI by a subset of researchers in the last two decades has caused a lot of recent research to be focused on implicit affordances, especially as they can be learnt by non-symbolic, mathematically well-defined, learning mechanisms, such as various kinds of statistical pattern recognition systems and reinforcement learning systems. However, perceiving, learning about, and using explicit affordances requires more than implicit affordances, and is a requirement for many uses of vision. I shall now try to explain these ideas in more detail.

### **3.8 Vision in action: transient, implicit vision**

Vision has many different roles. It can be used *transiently* in tight feedback loops that control actions such as grasping, or in transient triggers of ballistic actions such as throwing. That is probably how perception works in the vast majority of animals. Some animals, however, including humans, can also use perception in general, and vision in particular, to acquire less transient information that is stored in a form that may be useful for multiple subsequent uses, sometimes only for a short time, sometimes for longer. More importantly, if perceptual information is separated from the processes that use it, then it becomes possible for the same information to be put to different uses. By definition, this is impossible for implicit perceptual information.

Producing transient implicit information in action-control mechanisms is an important function of vision in humans and many other animals. For example, as you move a hand to pick up a mug, the perception of the changing gap between the hand and the mug forms part of a feedback control loop that enables the action to be performed fast, fluently, and accurately. As you run towards a gap in a wall, the complex relations between retinal inputs, head and eye movements, and currently sensed speed of running all feed into mechanisms controlling running actions and the direction of running. More ballistic uses of transient visual information occur in triggering protective blinking and saccades, in throwing, jumping and possibly ducking or dodging a fast moving object.

These control processes may use results of individual training and/or innate evolutionarily very old mechanisms and are shared with many other animals that use vision to control walking, running, jumping, biting, grasping, catching, hitting and even sexual responses. As suggested in Sloman (1989) the visual information in such cases may be fed directly into action control sub-systems, causing many sorts of changes, including acceleration and deceleration of moving parts of the body, switches of attention, increased alertness, and many visceral changes (often related to primitive emotions). Highly skilled performances of this sort need not make use of any explicit, re-usable, multi-purpose intermediate representation of distances, angles, speeds, locations, shapes, object-categories, or relationships of objects.

This does not imply that the same perceptual input always produces the same control signals, for at least two reasons (a) the mechanisms associating sensory inputs with triggered actions may be stochastic rather than deterministic, and (b) the action signals triggered may depend on both current sensory input and some internal state of the system. Internal and external sensor values could be combined to form an input-vector for a neural net, or both could be parts of conditions in condition-action rule-systems. (Either mechanism may be either stochastic or deterministic.)

None of this requires the sensory stimuli to have any use apart from their use in triggering the responses in the neural net or rule-system, possibly in combination with context vectors which modulate the effect of the stimuli. In that sense there is no need for any kind of action-neutral summary of what the input is, or what it implies about the environment. It is merely a conditional trigger, or partial trigger.

For animals whose perception is entirely like this (e.g. insects, or simpler organisms) the information that is implicitly represented cannot be translated into any language we would recognise. It need not, for instance be expressible as ‘that object is edible’, or ‘that berry is round’, or ‘that object is moving rapidly towards me’, even though information of that sort in a more sophisticated perceiver could lead to the same actions via a more indirect route.

In contrast, if you passively watch some event taking place and then later either report it, or mimic what you saw, or later move some object to prevent a similar event occurring, you must have stored information about the event in a form that makes the information re-usable in some other way than immediately reacting to it. In general, once a structure recording an object, event or process is created and stored, that structure can be given different uses, unlike information expressed *only* transiently as a pattern of activity within a mechanism that uses the information.

How explicit information is stored can vary enormously: it might be a persisting pattern of activation in a neural net, it may be a set of persisting weights in a neural net, it may be holographically combined with other persistent information in a distributed memory, it might be a persistent electromagnetic wave-form in a resonating mechanism, it might be stored in a complex molecule or in the frequency distribution of a set of molecules in a chemical soup, or in bit-patterns in a computer memory, or in virtual machine structures of some sort, such as virtual synapses in a virtual neural net, or an entry in a database. That is not an exhaustive list of possibilities for explicit information structures.

**[[JK: say something about the problems of access and the variety of types of solutions.] Various solutions in different parts of the architecture. Some neural nets provide a content-addressable memory with completion or interpolation capabilities. Hierarchical synthesis and structure sharing can be important. Context mechanisms are especially important at high levels of abstraction. Combining top-down and bottom up processing allows domain knowledge to speed up search, etc. etc. [Where should all this go?]]**

In organisms or machines that have both mechanisms that use transient implicit visual information ‘on-the-fly’ in controlling actions, and mechanisms that use explicit visual information, the same low-level sensory input mechanisms can drive both.

Transient visual information need not be very short-lived: if a certain pattern of visual input controls an extended action, then the implicit visual information can exist for the duration of the action, for instance the implicit visual information about the location of an object which controls the action of grasping the object.

The fact that explicit information is re-usable, moreover, does not imply anything about how

long it endures. Explicit information may also be very short-lived, e.g. because decay mechanisms are used, or because it is constantly being over-written by new input. Special recording mechanisms are needed if loss of information by constant over-writing is to be prevented.

### 3.8.1 Perceiving a static scene while eyes and head move

Sometimes preservation of explicit information may be a side-effect of another function. For instance if head movements or visual saccades are very frequent then the lowest level retinal stimuli are constantly being wiped out by new stimuli which may be discontinuously related to previous stimuli. Thus if the results of multiple head and eye movements are to be combined to construct a coherent information store about the current environment then there must be rapid transfer from retinal sensor buffers to some other data-structure that is not permanently in registration with the retina, whose mapping to the retina constantly changes and depends on how each saccade or head or body movement causes sampling of a different part of the scene (Trehub, 1991).

The structure containing integrated results of different movements will not represent retinal stimulation but something that does not change so rapidly: it could represent a 2-D array, which Gibson referred to as the 'optic array' defined by a certain viewpoint, which is sampled by multiple saccades. Or it might, after more complex processing, represent properties of visible surfaces, e.g. the 'intrinsic images' of Barrow and Tenenbaum (1978).

This enduring structure with information about the immediate environment will not have any simple relationship to the current pattern of retinal stimulation for it is the result of a *history* of many different patterns of retinal stimulation in the context of a history of known eye movements.

This information structure might be implicit if its only role is as a pattern of activation of a mechanism controlling a certain class of actions. This shows that implicit visual information need not be merely a pattern of retinal stimulation – it can be *derived* information.

If the information structure has multiple uses and can exist beyond its use in any particular action control process then it is an example of explicit information.

### 3.8.2 Precursors of 'episodic memory'

There is much research in psychology on episodic memories that record particular events, particular states of affairs, involving particular objects, people, and places. A key feature of such memory is that it presupposes an ontology in which there are enduring individuals, so that what one remembers can include *Fred*, or *Joe's car*, or a particular house, or city, or country. Such notions are also part of the requirement for a functioning visual system that can maintain information about the contents of a visual scene across saccades and other movements that disrupt the mapping between retinal image and what is perceived. Any such mechanism needs to have the notion of some object or place observed at time  $t_1$  being the same as an object or place observed at a later time  $t_2$ . This makes it possible also to include information about the relationships between objects observed at different times, and information about changes between saccades that are not changes in retinal patterns, but changes in the environment.

The need for an information structure that preserves information about the environment that is invariant across a possibly lengthy sequence of saccades, head movements and body movements producing many changes in retinal stimulation (sampling different parts of the optic array), has

much in common with the requirement of an animal that moves around some terrain to create an enduring information-structure indicating where things are, and how they are related to one another. The latter information-structure may take the form of a network of routes linking places in the environment. If those routes are treated as lying in a common metric space then it is possible, in principle, to glean information about directions and distances between points that have never been traversed on the same route. This amounts to having the ability to extrapolate from routes actually used to create the structure to routes implied by the structure (e.g. short cuts).<sup>7</sup>

It is interesting that although humans do have the abilities mentioned here they are notoriously inexact and unreliable. For instance if you look at a complex scene and then shut your eyes, and someone asks you to point at one of the objects or places previously seen you may not be able to point as accurately as if your eyes were open. Contrast the ability of marsh tits to bury large numbers of nuts in different locations, then later retrieve them. They even remember which locations no longer have nuts and do not go back after eating a nut. This is a highly specific ability that has evolved more than once in different species. There could be some general mechanism for doing this that is implementable in brains, but is too costly to develop unless there is a major benefit.

### 3.8.3 Implicit information and dynamical systems

In some anti-symbolic-AI circles it has become fashionable to claim that instead of symbol-manipulating systems dynamical systems are needed. Beer (1998) makes the slightly weaker claim that dynamical systems are ‘more fundamental’ than computational mechanisms. It seems that proponents of such claims are thinking mainly about types of perceptual processing which have here been called ‘implicit’. On this view all information is encoded in the states of the dynamical systems comprising sensors, motors, the body, the environment and possibly other internal dynamical systems. This may indeed be a good way to think about some functions of vision, e.g. controlling actions in tight feedback loops.

It is not so clear that the dynamical systems approach has anything useful to tell us about how someone can plan a future journey, how an engineer can design and build a complex machine, how someone looking at a machine can understand how it works, how mathematical formulae are read, how marsh tits retain information about locations of buried but so far uneaten nuts, how stories are composed, how stories are understood, and how many other characteristic human capabilities are explained.<sup>8</sup> Too often researchers find a good tool for a specific class of tasks, and then see it as the solution to everything. This has happened too often in the history of AI.

It is worth noting that this explicit/implicit distinction is a matter of how information is represented and used, and is not the same as the conscious/unconscious distinction. We can consciously perform actions using implicit information (e.g. using perceived spatial discrepancy to control grasping or catching), and we can unconsciously perceive explicitly represented structures and relations, for instance unconsciously perceiving relative sizes of objects, or unconsciously perceiving grammatical structures during language learning and comprehension, or unconsciously recording the location of an object so that you can later reach for it without looking.

---

<sup>7</sup>Compare the ability to point to a particular upstairs room from a downstairs room in a house permitting no direct (straight) route between the rooms.

<sup>8</sup>Compare the critique of dynamical systems theory in Sloman (1993)

### 3.9 Varieties of learning

Different forms of learning are required for implicit and explicit forms of perception. For instance, since implicit perception merely encodes what is seen in actions that are triggered, the only kind of learning that can happen is modification of the mappings from percept/context pairs to actions triggered. The context may include some internal state, such as a detected need, some record of the current situation or of recent history, or a partially completed action. Such contexts will make a difference to the action triggered by perception. If you are not hungry and food is plentiful it is pointless wasting energy chasing something edible that has been perceived.

The learning of *implicit* affordances is mainly a matter of changing probabilities of particular actions being triggered by different sensory inputs in different contexts (e.g. partly in response to different sensed needs, such as a need for water or a particular type of nourishment, or sensed sexual readiness). Reinforcement learning can achieve this.

In more sophisticated reactive systems it is also useful to use associative learning to change mechanisms allowing the current state to trigger not just actions, but also *predictions of results of actions* so that (a) anticipation becomes possible using feedback from the predictions and (b) errors can be detected by comparing predictions with subsequent sensory inputs.

All of this can be described as learning of ‘sensory-motor-contingencies’. In particular, at least two distinct sorts of contingencies can be associated with a particular perceptual input. The first are *action contingencies*: mappings between non-perceptual contexts (e.g. current sensed needs) and appropriate actions given a particular sensory input. The second are *prediction contingencies*: given this particular sensory input, what predicted outcomes should be associated with different actions? In the case of implicit perception, both action contingencies and prediction contingencies are implicitly encoded in the mechanisms that trigger the actions or generate the predictions. They need not be explicit premises or rules that can be used to make different kinds of inferences from different combinations of premisses.

Such learning of contingencies can be done in various ways, including adjusting weights in neural networks or modifying strengths of condition-action rules in a ruleset. A neural net can be seen as simply a particularly simple kind of condition-action rule-system, where many rules are run in parallel, where all conditions and action signals are restricted to weighted combinations of numerical values, where rules can share conditions and can share actions, and where the only structural side-effects are changes in the weights.<sup>9</sup>

In more sophisticated cases there may also be a need to learn new concepts, i.e. new ways of partitioning the space of low-level sensory input arrays, or input array sequences. However there is no unique right way to do the partitioning: it will depend on the animal’s or robot’s needs, goals, and action capabilities. In more complex cases the learning may require growing new sub-nets or creating new rule-sets.

We shall later see that there are some very subtle and sophisticated kinds of visual affordances, and that learning them requires more complex mechanisms and forms of representation than those required for segmenting and classifying objects.

---

<sup>9</sup>Compare the ‘causal diversity’ matrix in Minsky (1992)

### 3.10 Explicit associative perceptual information

In contrast with the transient use of implicit visual information to control action, there is also the use of vision to acquire information that is represented explicitly in some structure that can be stored and then used, potentially in different ways, later on. If you see a big ripe apple in the tree in your garden you can later use the information thus gained when you are hungry, when you want something to throw at an intruder, when you wish to annoy the owner of the tree, when you want to impress a visitor, when you tell someone where to find the apple, when you paint a picture of it, and in myriad other ways.

However there is a difference between seeing the apple and seeing something else that informs you that there is or was an apple present. For instance an expert may see from the curvature of a branch that there is an apple weighing it down, even though the apple is out of sight behind a wall. This uses a learnt association between apples and curvature. Even a non-expert could see someone in the distance pick something off an apple tree and start eating it: and would describe this as seeing the person eat an apple, even though the apple itself was not seen because it was occluded first by leaves, then the hand holding it then something blocking the view of part of the person's face.

We can contrast those cases with what could be called 'direct perception', where features in the sensory field (or Gibson's optic array) are caused by parts of the apple reflecting light or obscuring light reflected from other objects.

An intermediate case is seeing an object part of which is occluded: we normally say 'I saw Fred at the far side of the table', not 'I saw the upper half of Fred at the far side of the table'. The inference that the whole of a partly visible object is present is not typically made by a central general-purpose reasoner — it uses some sort of dedicated, fast, automatic, highly trained mechanism required for rapid perception of complex scenes.<sup>10</sup> But in some sense it is associative, based on the general information that half humans do not talk, pick up knives and forks, etc.

For humans, reading is an important type of associative, indirect perception, e.g. reading a note from a trusted source, saying that there is a ripe apple a third of the way up the south side of the second tree from the front door. When we read a story we are using our eyes to learn what happened to all sorts of people and things in the story, and it does not matter whether it is a true story about real people and places, or an invention about real people, or a pure fantasy about people and places that never existed. Moreover, unlike seeing, when we read, the information gained about the people, places and events has no structural relation to the visual input: you can read a sentence near the bottom of the page describing something happening at the top of a tower whose base was described at the top of the page, or on another page. The description of the middle portion of the tower does not need to come between descriptions of top and of bottom. The description of what happened in the middle of a certain day need not come between descriptions of what happened in the morning and evening of that day. You can read about someone looking down from a bridge, without getting any information about how high the bridge was, whether the person was male or female, whether the person was wearing clothes, or what sort, which way the person is facing and so on, whereas *seeing* someone on a bridge not too far away provides a lot more specific information, derived from the structure of the optic array.

The relations between symbols on a printed page and what they describe is not totally arbitrary, for without some systematic relation we could not read new sentences about new events: novels

---

<sup>10</sup>Contrast the role of a general purpose associative reasoner in considering the possibility that what is seen is some sort of illusion, e.g. when what is seen is on a magician's stage.

would be impossible. But the relationship is very abstract, comparable to the relationship between a mathematical formula and what it denotes (Sloman, 1971).

The ability of vision to provide such vivid information through reading is closely related to the ability of speech to provide the same information. Both need to be explained as part of a general theory of perception, but for now we shall ignore perception used in the service of language comprehension. That sort of perception is a late addition to other sorts of perception that are evolutionarily older and shared among far more animals. The ability to read and to understand speech may build on other non-linguistic uses of perception to gain information through learnt associations, e.g. seeing a bent branch as indication of fruit at the end, seeing dark clouds as precursors of rain, seeing tracks in the soil as an indication of presence of prey, or predator, in the vicinity, or seeing the colour of fruit as an indicator of taste.

### **3.11 Non-transient direct perceptual information**

In contrast with the previous cases, where seeing involves acquiring transient, implicit information represented only temporarily in control states, or seeing involves use of a learnt association between visual information and something not directly visible, there is a third case namely seeing that provides non-transient, re-usable information about things, properties, relationships and processes that are in view, so that the information acquired is structurally related to what is seen, as well as being optically caused by it.

The question for us is: what does such seeing amount to? The answer implicit in much research is roughly the answer given by David Marr, namely that seeing involves using our eyes to acquire information about textures, colours and orientations of surfaces and the location, structure and relationships of many objects that are in view. Our answer is that this leaves out some of the most important functions of vision, namely perception of non-physical aspects of other agents and perception not just of existing structures but of affordances, which are inherently concerned with what *might* happen or exist. Perception of non-physical aspects of other agents includes things like seeing a face as happy or sad, seeing someone as looking intently, or as trying hard to do something. This is in some ways like the associative non-transient information discussed earlier, though the phenomenology is different: it feels more like seeing than like inferring or interpreting. I shall return to that point later.

#### **3.11.1 Perceiving structures and perceiving affordances**

For now, let us leave aside perception of mental states and processes in other agents, and focus on perception of scenes involving only physical objects in various relationships. In such cases there are (at least) two importantly different aspects to the task, e.g. of understanding a picture, diagram or movie, or visually perceiving the environment. The first aspect involves *perceiving the structures that are present* and the second involves *perceiving affordances*. This includes perceiving causal relationships and especially relationships to possible actions. This means perceiving what can and cannot be done in the situation and what the consequences of possible actions would be. However, this depends crucially on the fact that the perceiver is an agent in the world perceived even when there is no agent perceived in the world.

So perceiving affordances involves gaining information not about features that are intrinsic to

the physical objects, properties, relationships in the environment but rather gaining information that relates to the perceiver's possible or actual goals, actions and capabilities. This might misleadingly be described as perceiving 'subjective aspects' of objective situations. More accurately it is a matter of perceiving *relational* properties of things in the environment where the perceiver or some aspect of the perceiver forms a term in the relation.

So, whereas the main focus of work on perception so far has been concerned extracting information about what exists in the scene, of the sorts listed by Marr, perceiving explicit affordances involves perceiving more than that: it requires seeing what *might* exist or happen, and what *cannot*. Those possibilities that are relevant to achieving an agent's goals can be labelled *positive* affordances, and those which impede actions or goals are *negative* affordances. The very same object may play a role in both positive and negative affordances. A table can allow objects to be placed where they can easily be picked up again, a positive affordance, while getting in the way of direct motion to the door, a negative affordance. Which set of possibilities and constraints constitute the affordances in the scene is relative to the perceiver. However, more sophisticated affordance-perceivers can take account of the goals and capabilities of others, and perceive other-related affordances. E.g. a parent or nursemaid looking after an active child has to be very aware of affordances relative to the child, for instance. I suspect humans are not the only animals able to do that.

These points, and related points in my previous papers on, vision were inspired by Gibson's notion of 'affordance' in Gibson (1986) – originally published in 1979 – though I doubt that he would agree with everything said here.

Affordances, as construed here, involve possibilities for action and constraints on possible actions, so that they include information about processes and objects that do not exist but might exist. The affordances in a scene are all relevant to particular agents, and the same environment can have different affordances for different agents. In particular the affordances for an agent A in a particular situation will depend on

- the goals A can have (which A need not necessarily have at the time)
- the actions A can perform

This implies that as an agent develops, perhaps changing size, acquiring new skills, and developing new types of goals, the affordances that are relevant to that agent will change.

Perceiving affordances is essential for many types of human and animal competence. Later I discuss the objection that this is not a case of seeing but inferring or deducing. In any case, it involves an understanding of spatial structures and processes far beyond what current AI systems seem to be capable of.

The contents of visual perception vary in a number of different ways, including the scale of perceived objects, the degree of abstraction or interpretation, and the differences between seeing existing structures and seeing other things such as causal and functional relationships, and also affordances. The next section elaborates on some aspects of the variety of structures that can be seen, and later sections illustrate perception of non-geometrical, non-physical aspects of some objects as well as perception of affordances.

### 3.12 Perceiving perception: self-awareness

It is often assumed (sometimes unconsciously) that all perception has to be conscious, e.g. it is impossible to perceive anything without being aware that you are perceiving it. This is wrong. That is wrong, and why it is wrong, becomes clear if we understand the architectures within which can play a role.

In the last two decades research in AI on forms of representation and algorithms or mechanisms has been increasingly supplemented by work on architectures within which different kinds of functionality may be combined in an integrated system.

Refer to work on meta-management and the role of meta-semantic capabilities, including the requirements for first-order, second-order and third-order ontologies.

Choose from:

Beaudoin (1994) Wright et al. (1996)

Chapters 6 and 10 of Sloman (1978) Sloman (1990)

Sloman (1993) Sloman (1994) Sloman (1996b)

Minsky (1987)

Sloman (1997) Sloman (2000a) Sloman (2002b) Sloman and Logan (2000)

Sloman (2000b)

Sloman (2001a) Sloman and Chrisley (2003) Sloman and Chrisley (2004)

Sloman and Scheutz (2001)

Minsky et al. (2004) Sloman (1989)

Explain how meta-management can include implicit and explicit (introspective) perceptual information. Some of the meta-management system's meta-semantic information is about what is being perceived.

### 3.13 Attention

An important topic that has not so far been mentioned is attention. It is perhaps arguable that the need for attention is a by-product of resource limits. If an organism had sensors placed everywhere in its environment and had unlimited power to process sensory inputs in all possible ways, there would be no need for attention. As things are, choices have to be made. For instance, eyes cannot look in all directions at once, and even within a perceived part of the visual scene not all parts can be seen with the maximum resolution because of the structure of the retina. So body and head movements and both tracking motions and saccadic motions of eyes can select what to process and with how much detailed information.

At a more abstract level the same information can be processed in different ways. E.g. trying to detect small movements requires different processing from trying to see where you dropped a berry. Trying to estimate the distance of something looking at you requires different processing from trying to determine whether it has noticed you, or whether it is angry. Judging the distance in order to throw a rock is different from judging the distance in preparation for a jump.

A full account of attention would require an analysis of all the kinds of *choices* that need to be made in the many sub-systems operating concurrently on tasks with varying levels of abstraction and varying levels of importance. Notice that this is very different from the old idea that attention is primarily a matter of filtering sensory input. That idea is concerned with choosing *what* to process, but the kind of agent we are discussing also has to choose *how* to process information, on the basis of what its current needs are.

Yet another issue concerns the common assumption that attending to something involves being conscious of it. This may be true of the selections made in a meta-management layer of a multi-layer architecture, but many more choices are constantly being made in other sub-systems and they all involve the generalised notion of attention as selection. For example, we are typically ignorant of our own visual saccades which constantly switch attention in low level visual processing mechanisms.

The question of how choices are made will be mentioned briefly later in connection with the problem of which affordances are relevant in situations where typically there are far more affordances than can be processed. This is one of many *control* problems in intelligent systems.

### **3.14 Aesthetic aspects of perception**

There are many ways in which the discussion so far is incomplete. One important gap is the lack of anything about affective aspects of perception – e.g. what does it mean for a visual experience to be pleasant or unpleasant, aesthetically and in other ways.

More to be added. Doing this properly requires an account of affective states and processes: what they are, why they are needed, why they sometimes arise as dysfunctional side-effects of other things, how they are implemented, etc. Some relevant background ideas can be found in Sloman et al. (2004) (to be included in a volume entitled *Who Needs Emotions?: The Brain Meets the Machine* edited by Arbib and Fellous (OUP 2004).

(Compare A. Sloman, M. Croucher, (1981,) ‘Why robots will have emotions’, *Proc 7th Int. Joint Conference on AI*, pp. 197–202)

### **3.15 Perceiver-centred vs vicarious perception**

It is worth noting that whereas all forms of perceptual information discussed so far are perceiver-centred insofar as they are concerned with what a perceiver can do or predict or learn, there are some animals, including humans, that are able to perceive what is happening or could happen to them, but also what is happening to another or could happen to some other object or agent. This covers a wide range of cases and a full survey is not possible here. But it is worth noting that in altricial species, whose young are born relatively incompetent and learn by being active as they grow and develop, it can be useful for adults to perceive opportunities, obstacles and possible dangers confronting their young. This covers such diverse cases as noticing that a child is approaching a deep pond, noticing that a child will find it easier to pick up a cup if it is rotated with the handle on the child’s right, or noticing that the child will find it easier to avoid spilling the contents if the cup is only half full, and many, many more.

In relatively unintelligent animals such vicarious perception may use only implicit information,

for instance if the approach of a possible predator directly triggers some innate mechanism that causes the mother to move between the infant and the predator, or triggers decoy behaviour. The mother need not have any understanding of the potential consequences of the approach of the other animal, nor the consequences of its own behaviour. If the parent is able, in addition, to represent the impending approach explicitly, it may be able to reason about the consequences, investigate alternative remedial forms of action, evaluate them in relation to preventing undesirable consequences, and use all that to select an appropriate action, which may involve aggression, diversion, decoy, protection, etc. depending on circumstances.

In some cases vicarious perception can lead to learning. If an inexperienced animal or machine observes something good or bad that is about to happen to another, then sees the other take advantage of it, or take preventive measures, that perception can lead to future actions in the perceiver that have similar effects. The ability to learn in this way requires solution of many problems of representation which are ignored by those who assume that imitation is an explanatory mechanism which needs no further analysis, or that so-called ‘mirror neurons’ suffice to explain imitative learning. Putting the ability to imitate in the context of vicarious perception, including perception of positive and negative affordances for another individual can help us appreciate better what the problems are.

### **3.16 Summary so far**

Figure 1 summarises the varieties of types of perceptual information distinguished so far. Further distinctions need to be made between different kinds of *uses* of the information. For example, perceptual information used to guide action may be used either in a process of analysis, planning, and careful decision making, or it may be used in ‘alarm mechanisms’, which use fast, and possibly less reliable, pattern matching mechanisms that are capable of overriding other processes and globally redirecting processes either within a sub-system or in the whole organism. Such mechanisms are often needed in complex control systems and have been used by engineers for a long time. In Sloman and Croucher (1981) it was shown how side effects of such mechanisms could account for some aspects of human emotions. However, alarm mechanisms are not the topic of this paper and will not be discussed further, though they are interesting and important within a general theory of architectures for intelligent organisms or machines, as explained in Sloman et al. (2004).

## **4 Perceiving structures, at different levels of abstraction**

In a 2-D graphical display there will be two-dimensional features such as colours, intensities and textures, optical flow, gradients of various kinds (intensity, texture, colour, flow), boundaries where the gradients are sharp, spatial gradients in intensity, colour and texture, and flow gradients and boundaries. All of these will be located in a manner that can be specified using co-ordinates or other representations of image points (and times), or contiguous sets of such co-ordinates making up regions. In computer vision systems these features are often arranged in a rectangular 2-D array, though other configurations are possible including hexagonal arrays, collections of concentric rings of variable resolution receptive fields, linked networks of image locations, etc.

In a computer generating such a display there may be several representations including both

## KINDS OF VISUAL INFORMATION

**1. Implicit perceived information**, expressed transiently in the current state of a dynamical system, e.g. states of (deterministic or stochastic) neural nets or active hormonal systems mapping combinations of internal and external sensor inputs into some kind of output vector, e.g.

- (a) *in action controllers*, including
  - i. implicitly sensed external action-contingencies
  - ii. implicitly sensed internal action-contingencies
- (b) *in predictors*, including
  - i. implicitly sensed prediction-contingencies for external actions
  - ii. implicitly sensed prediction-contingencies for internal actions

**2. Explicit perceived information**, (‘direct’ or ‘associative’) using dedicated visual mechanisms in registration with sensory arrays, expressed in some less transient structure or state capable of being combined with different kinds of information, and/or used for different purposes:

- (a) *in reactive subsystems* (with no ‘what if’ representational or reasoning capabilities)
- (b) *in deliberative mechanisms* (able to represent and reason about hypothetical past, present or future possibilities)
- (c) *in meta-management systems* (using meta-semantic capabilities to monitor, categorise, evaluate, learn from, and to some extent control information-processing sub-systems)

where the information may be about perceived

- *objects, properties, structures, relations, processes*, i.e. things that *exist* in the current environment (using first-second, or third-order ontologies)
- *affordances*, i.e. things that *can or cannot exist* or be done in the current environment

where complex information may be expressed as

- *vectors/arrays of values* interpreted in parallel (e.g. by matrix operations, convolution mechanisms, using neural nets, chemical soups or other statistical combinatorics)
- *hierarchical/recursive structures* interpreted using compositional (essentially sequential) semantics, using one or more of the following modes of representation
  - logical
  - analogical
  - procedural
  - others, including mixtures

and where the processing from sensory input to percept may be

- *flat* i.e. a single step from low-level feature detectors to information about what is perceived
- *layered* i.e. making use of intermediate layers of interpretation (as in hierarchical synthesis) using either a *single pipeline* or *multiple routes* performing different kinds of analysis in parallel – e.g. texture, colour, surface orientation, optical flow, etc.,

**3. Derived, ‘associative’ information**, produced from perceptual information by a general-purpose reasoning or inference mechanism that is not dedicated to perceptual processing and does not produce results that are ‘in registration’ with sensory input arrays.

Figure 1: *Varieties of perceptual information - a high level partial overview.*

[DRAFT: The table does not include all distinctions in the text.]

some sort of high level description and a low level specification of patterns of behaviour of streams of electrons generating illuminated points on the screen. Even if there is a systematic mapping between the low level specification used to generate the display and a low level sensory configuration in the perceiver's visual system, the computer will generally have no notion of the higher level structures perceived by a human looking at the screen.

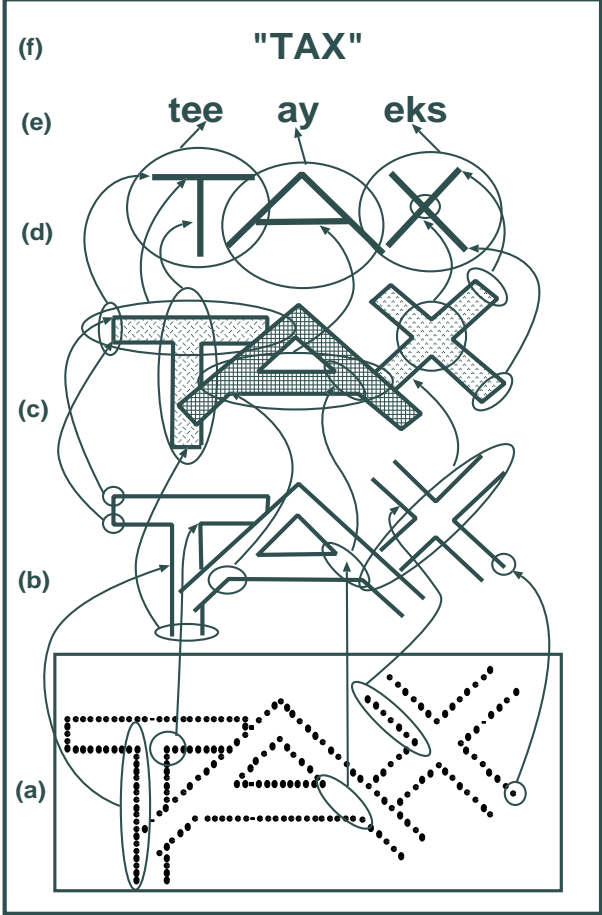


Figure 2: Layers of interpretation of a 2-D dot pattern (based on Sloman (1978), chapter 9).

### 4.1 Larger 2-D structures and more abstract 2-D structures

Within the sensory array, a human (and some other animals?) can detect higher level 2-D structures that are made up either of *collections* of these structures or *more abstract interpretations* of the features and collections of features.

#### 4.1.1 Agglomerations in 2-D fields

The collections, or agglomerations, of features include such things as continuous groups of intensity discontinuities, texture discontinuities, optical flow discontinuities, or other small features, forming

lines of various sorts, or regions made of continuous sets of points which share some feature or vary only a little in respect of that sort of feature, or some feature-vector. Examples are groups of dots in Figure 2 taken from chapter 9<sup>11</sup> of Sloman (1978). More complex hierarchical structures are formed from groups of these aggregation structures, for instance a junction where two or more linear features meet, or a collection of regions arranged in some regular pattern. In a changing display there will be more complex 3-D (two spatial dimensions and one time dimension) structures in the changing display.

#### 4.1.2 Abstract features perceived in 2-D fields

More abstract 2-D structures, can be seen in (or some would say hallucinated onto) the initial configuration of features, such as continuous “infinitely thin” lines, line-junctions, enclosed regions, or more complex objects composed of such things. These include the sorts of entities characterised in Euclidean geometry and are subtly different from the mere groupings of sensory features. For instance a perceived intensity discontinuity with a certain degree of fuzziness and minor variations from straightness can be interpreted as representing an infinitely thin, perfectly straight line segment, which could not be directly sensed. (It is not clear whether other animals can perceive instances of Euclidean geometrical objects in their sensed 2-D or 3-D structures, nor whether and how this ability develops during infancy in humans.)

Additional abstract entities and configurations of entities related to particular domains of expertise, e.g. reading text, can also be seen in (hallucinated onto) visually sensed configurations of features. Figure 2 indicates crudely how a 2-D pattern of dots can be interpreted in terms of several layers of 2-D agglomerative structures, in addition to more abstract structures such as sequences of letters forming words where the letters are abstract 2-D entities made up of abstract strokes, stroke-junctions, etc. Some of the intermediate interpretations may include a depth ordering, so that certain objects are seen as on top of others, e.g. the interpretation labelled (d).

In pictures where the letters are fairly clearly separated it would be possible for template-based pattern recognizers, or neural net mechanisms, to go straight from (a) to (e) just by detecting high level features in groups of dots. However if the letters are more jumbled and there are spurious dots, as in Figure 3, segmenting the letters without going through the intermediate levels can be very difficult, yet people degrade gracefully in their ability to detect the words as the messiness increases, as did the Popeye program reported in Sloman (1978). Although this is an artificially generated image, it illustrates aspects of naturally occurring problems of vision in humans and other animals, such as cluttered scenes with variable amounts and kinds of occlusion – whether caused by solid objects, fences, dirty glass, shrubbery, falling snow, etc. In addition there are well known problems of seeing things under conditions of variable lighting, speckled shadows, motion of whole objects or relative motion of parts of flexible objects, and often in conditions where despite the difficulties, perceptual decisions have to be made quickly e.g. because of the risk of missing opportunities for lunch or escape from being lunch for someone else, or the risk of grabbing the wrong branch to break your fall from a treetop.

Real seeing is often very much harder than the tasks most artificial vision systems can perform at present. It is also often much harder than many of the tasks presented by psychologists to subjects

---

<sup>11</sup>Also available as <http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html>

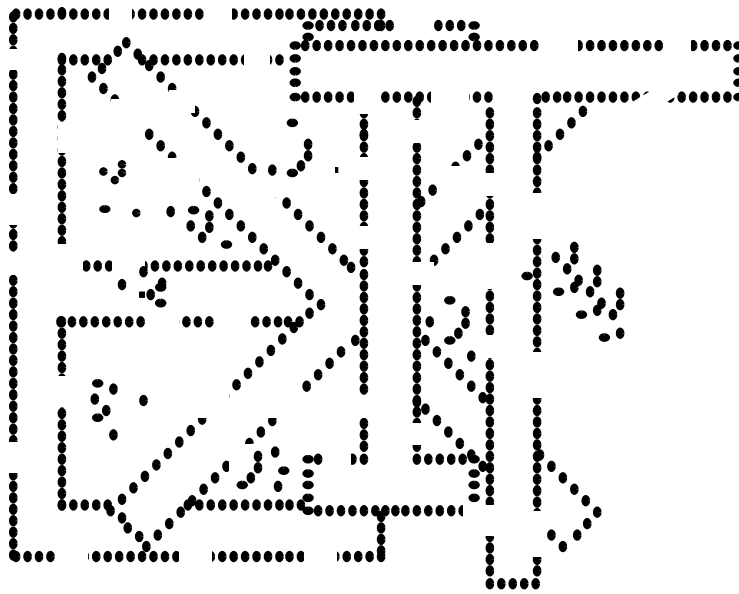


Figure 3: *Illustration of graceful degradation in human vision.*  
(based on Sloman (1978), chapter 9)

in vision laboratories – tasks selected for suitability for repeatable laboratory experiments.

Such graceful degradation in the face of various kinds of messiness, appears to depend on our perceptual mechanisms being able to operate at different levels of abstraction in parallel, with information flowing bottom up and top down, as postulated in the chapter referenced. The idea of a mixture of top-down and bottom up perception is very old. E.g. it was labelled ‘analysis by synthesis’ in the 1960s, in Neisser (1967) and has sometimes been described as ‘hierarchical synthesis’. Similar ideas were explored (and implemented on painfully slow computers) in some speech understanding systems in the 1970s.

Part of the task in understanding images like Figure 3 is understanding the difference between dots or lines that are missing because of random noise and dots that are missing because they are hidden by something. The latter requires the machine to use a richer ontology, including notions like “nearer”, or “behind”, for interpreting its sensory information. With a suitable ontology available the interpretation process in the perceiver can ‘postulate’ something invisible that reduces the noisiness of the picture, e.g. by ‘completing’ a partially invisible object. Later we shall introduce more subtle requirements for a perceiver’s ontology.

## 4.2 Geometric and non-geometric 3-D spatial interpretations

So far we have illustrated percepts that vary in scale (for instance the number of component dots, or the sizes and number of component lines or line groups) and also vary in the kind of ontology used for the interpretation of sensory data, even while remaining very close to 2-D image structure.

Humans, many animals, and robots also need to be able to use vision to acquire information about a three-dimensional spatial environment, where objects can vary in distance from the viewer, where surfaces can have different orientations, or can be curved, and where the 2-D appearance of an object can vary for other reasons than noise.

In a three-dimensional world, there are far more complex structures and relationships that can be detected in and ‘hallucinated’<sup>12</sup>. onto images. For instance one kind of complexity, evident in the “flat plate” interpretation of the dot array shown in level (d) of Figure 2, includes objects having parts that are not visible, either because they are obscured by nearer objects or because they are on the far side of a perceived object. Another example is self-occlusion in a cube with several faces not visible, or partial occlusion of a table-leg by the top of the table, as illustrated in Figure 4.

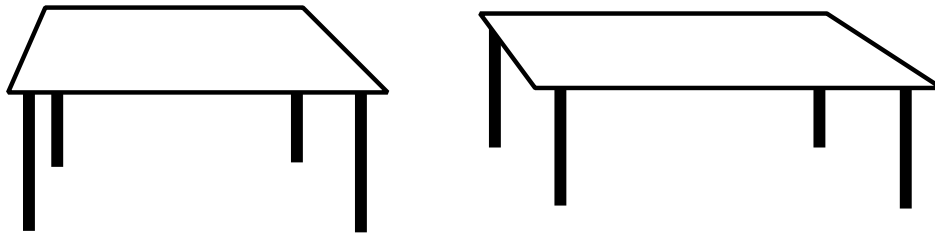


Figure 4: *Viewpoint-dependent shape and occlusion.*

This raises questions about the kind of information structure built up in the perceiver. If an object is perceived as hiding a part of another object whose full structure is known (like the laminas in (d) of Figure 2) the conventional and obvious way to design a visual system would lead to a mechanism that builds a representation of the complete sub-object, including invisible parts such as edges and surfaces continuing behind the obscuring object – something like the sort of representation that might be used in a computer-aided design package for complex 3-D objects. This more complete representation would somehow have to be linked to the less complete representation depicting only visible portions. This could be done in many ways including pointers in one or both directions between image locations and portions of the representation of 3-D structure, or by labelling parts of the complete structure to indicate their visibility.

Where two linked representations are created, each would be useful for some purposes and not others. For instance the image structure with its 2-D relationships that change with viewpoint (as in Figure 4) would be relevant to an artist painting a picture of the scene, or a steersman guiding a boat into harbour by keeping two points in the scene aligned in the 2-D image space. The representation of the 3-D relationships would be useful for answering questions about functions of the object, how it would look from a different viewpoint, how a hand should be shaped to grasp it, or, in the case of the steersman, how long it will take to get into the harbour. (A more complete account of the uses of each type of representation leads into the discussion of affordances, below.)

It is not necessarily the case that the perceiver knows or can infer the structure, colour, texture and other features of the invisible portions of perceived objects. For instance if Figure 2) were changed to contain a “dotty-lamina” representation of an Arabic sentence, a person who did not know the relevant alphabet would not know the shapes of hidden portions. This could also be true for a familiar alphabet using an unusual font.

In that case, the perceiver may need to be able to construct representations that include information about what information is missing. It may also be useful for the representation to indicate what to do in order to answer specific questions about invisible portions of the scene – e.g. how to change viewpoint, or how to rotate the object.

---

<sup>12</sup>Von Helmholtz described seeing as “controlled hallucination”

The problem of how we see partially obscured objects is multiplied a thousand-fold when we gaze around a cluttered room with large numbers of objects of different shapes and sizes, many of which are partly hidden. For dealing with the problem of finding out about hidden parts of objects there could be rules or other mechanisms “attached” to the representations of visible parts specifying what to expect if you move your head so as to see more of the object. That might reduce the problems of searching for an appropriate action.

This is a proposal made by Minsky long ago Minsky (1978) in which 3-D objects are represented with information making it easy to determine how visibility of parts will change with changing viewpoint: “the effects of important actions are mirrored by transformations between the frames of a system”. He was thinking of a symbolic data-structure but similar general mechanisms might be implemented in a network of neural networks whose activity levels can change in response to inputs representing possible actions.

There are very many different ways in which the perception of something as continuing behind another might be implemented, though they vary in how well they explain the *phenomenology*, i.e. the details of the experience of seeing the right hand edge of the “T” shaped lamina as continuing behind the obscuring part of the “A” shaped lamina.

Later we shall see that Minsky’s 1978 proposal for providing information about what to do in order to acquire more information is a special case of a more general strategy for seeing affordances.

### 4.3 Combining structural and procedural information

The preceding discussion implies that in a cluttered scene, information about invisible portions of objects may be represented either (a) *concretely* with hypothesised details derived, perhaps using default reasoning, from prior knowledge (tables have straight edges) or by interpolating (assuming a straight connection between two collinear visible edge-fragments) or (b) *abstractly*, perhaps using procedural representations that hold information about how to generate hypotheses about the invisible bits and how to generate actions to test them.

For a complex object such as a table, or a tree, or a complex scene containing many objects, there will be many questions that could be asked about different portions of the object or the scene, and many actions that could be performed in or on different portions of the object or scene. For instance, a table can be moved further or brought nearer, rotated in a horizontal plane about a vertical axis or rotated about many different horizontal axes, or lifted in various ways (e.g. enough to allow a wedge to go under one of the legs, or lifted to enable the whole table to be carried, etc.) These different actions would require different parts of the table top or its legs to be grasped and moved in different ways.

Actions required in order to fill information gaps are a special case: actions can, of course, meet other needs. Understanding a scene can include having information about a host of potentially useful actions that are possible in that scene, and which serve many different sorts of purposes. For instance moving an object sideways may sometimes be useful for providing information about more remote objects and sometimes useful for enlarging a narrow passage through which one can move. It is useful to know about harmful consequences as well as useful ones.

Thus we have the idea of large numbers of possible actions “attached” to various objects or portions of objects and large numbers of possibly useful, or in some cases, harmful, consequences of those actions. There are different ways in which this information can be acquired, stored,

accessed and used, in different architectures.

#### 4.4 Proto-deliberative reactive perceivers

In an organism with a purely reactive information processing architecture, all of that information may be encoded in condition-action rules (which might be expressed in a neural net or some other mechanism) where the conditions include both perceived external conditions and also internal states, such as goal states, or states including previously acquired information. In that case when the conditions are satisfied a corresponding action may be automatically triggered: e.g. jumping, crouching, moving sideways, retreating, rushing forward, creeping forward, grasping, biting, etc. In general, however, any particular context will activate or inhibit a variety of possible actions, perhaps with different strengths of activation or inhibition, and some selection mechanism based on those strengths may determine which of the possible actions is performed.

As this is a primitive version of what deliberative systems do we can call this a *proto-deliberative* system. Proto-deliberative mechanisms must be wide-spread in organisms with purely reactive architectures. By contrast a deliberative system is able to create, compare and evaluate novel descriptions of possible futures by combining elements such as possible actions, and consequences of the actions, for variable numbers of steps ahead. This requires not only the ability to perceive possibilities inherent in the current scene, but also to perceive chained sequences of possibilities that would be enabled by selection of prior possibilities: the kind of ‘what if’ reasoning many AI planning programs do and also many humans planning complex future activities, such as getting to a conference in another country.

#### 4.5 Deliberative perceivers

The architectural requirements for effective proto-deliberative and deliberative systems are different, as we have indicated elsewhere, e.g. Sloman (2001a, 2002a). In particular, proto-deliberative mechanisms require prior evolution to have produced a linked network of condition-action nodes, though some kinds of individual learning (e.g. reinforcement learning) could alter the behaviour of such a network. In contrast a deliberative system can use search mechanisms to create novel complex action sequences, though not all deliberative systems have the same power.

A deliberative system that learns about its environment will need to acquire generalised reusable information that is in a form that is appropriate for exploring branching futures. This requires perceptual mechanisms that can “chunk” objects and processes in the environment so that they can be recognize as components for which generalisations hold, for instance the generalization that if an object is resting on another then moving the lower object will move the upper object. The ability to use this generalisation requires the ability to perceive the ‘resting on’ relationship, which is not the same as the ‘above’ relationship since one thing may be immediately above another but not resting on it. The ‘resting on’ relationship involves causal connections not implied by the ‘above’ relationship.

The ability to recognize properties and relationships that are relevant to possible actions may be implemented in very different ways. At one extreme there is the use of general-purpose reasoning mechanisms that derive consequences from low level information provided by sensors. At another extreme the perceptual mechanisms can develop (through evolution, or possibly through training)

new special-purpose capabilities for detecting more and more abstract categories and relationships in the environment, as sketched in chapter 6 of Sloman (1978) and in Sloman (1989). Whether this is possible will depend on whether there are sensory features that provide reliable cues to the more abstract properties. In the 1978 book such an environment is referred to as a “cognitively friendly environment”.

For someone who does not often do such things, that sort of practical information might be derivable using general reasoning about shape, forces, the causal consequences of various movements of hands and arms, etc. Or it might be derivable by trial and error, which appears to be what happens much of the time in infant learning, before the child acquires an understanding of spatial structure and motion.

For animal species whose members do not have the ability to reason about how to perform such tasks, and do not have time in the lifetime of individuals to use trial and error learning, it could be the case that evolution, over millions of years, generates the information, and stores it in genes, so that it can be transmitted to individuals. (Examples might be nest-building in birds and insects.) This requires that appropriate environments were encountered in the past that made it worthwhile for a species to develop the special information. It also requires that the variety of situations and actions is not so great that the information cannot be encoded in genes or cannot be encoded in the individual’s nervous system. Many varieties of insects seem to have found types of condition action information that have proved consistently useful over millions of years, and which can be largely transmitted genetically (though genetic information can often be compressed because the environment itself helps to control behaviour).

Besides the time required to acquire information whether by individual learning or by evolutionary development, there is another way in which time is important: producing actions in response to detected conditions will take time and if things move quickly in the environment, the triggering of actions may have to be very fast, perhaps too fast for deliberative mechanisms that reason using general-purpose mechanisms and representations.

## **4.6 Storing information for rapid access**

Even for someone who has sufficient understanding of the possibilities for action and motion in 3-D structures to work out what needs to be done to achieve various desirable ends or to avoid undesirable states, there may be a problem of speed. If a certain sort of task is performed often, it may be worth saving the results of previous derivations instead of repeating them whenever required. For instance someone who often has to move furniture around, such as a house cleaner, it may be worth storing re-usable information about which movements can be applied to parts of a table and what the consequences will be, instead of deriving it on every occasion. This is a requirement for a wide range of human skills, including athletic skills and skills in musical performance. For animals that have to be able to move quickly through tree-tops, such as monkeys and chimpanzees, there may simply not be enough time to work out exactly how best to act at every moment. Likewise hunting animals that need to capture prey that adopt skilful evasive actions.

How such re-usable information is stored can make a considerable difference to how accessible it is and how much searching is required in order to select it for a particular task. Perhaps this could be done in some cases by somehow attaching condition-action rules to representations of different parts of the object (e.g. “if this leg is grasped here and pushed in that direction the table will rotate

counter-clockwise in a horizontal plane”, “if this leg is grasped here and moved vertically upwards the surface of the table will tilt to the left”, etc.).<sup>13</sup>

The discussion so far shows that the products of visual perception may be of many different kinds, differing in scale, in the ontology used, in the type of information (e.g. structural, causal or procedural), in the manner of derivation, and in the form in which it is represented. Understanding more precisely how all this works is a prerequisite for understanding human and animal vision, for building robots with human-like visual abilities, and also for building machines that communicate effectively with humans using graphical displays that are understood equally well by the user and the machine.

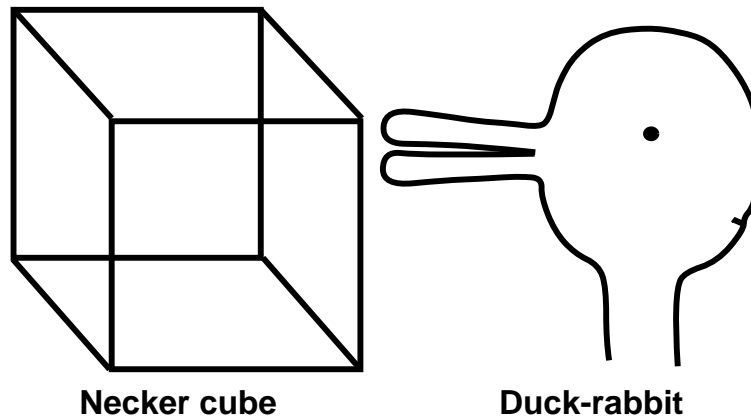


Figure 5: *Examples of geometrical and non-geometrical ambiguity.*

## 5 What is seen: clues from different sorts of ambiguity

It is common to assume that visual perception functions primarily or exclusively to provide information about physical properties of the environment, e.g. shape, motion, colour, texture, etc. But we have already seen that in complex scenes there may be more abstract percepts concerning for example how to move in order to see more of an object, or what to do to achieve other goals, generalising Gibson’s notion of “affordance” Gibson (1986).

### 5.1 Perceiving non-geometrical features, states, processes

By analysing and comparing the precise nature of various kinds of visually ambiguous pictures we can gain more insight into what is seen in those pictures. For example Figure 5 shows an image, the Necker cube, whose interpretation flips between two 3-D structures, both cube-shaped, but differing in which face is nearer or further, higher or lower, and in the slopes of the edges (e.g. sloping down and away from the viewer or up and away). It also shows an image, the duck-rabbit, which when it flips displays no *geometrical* change, only more subtle changes relating to which animal parts are

---

<sup>13</sup>For a recent proposal regarding the association of information about actions with the objects to which the actions might be applied see Steedman (2002). A related proposal was also made in Pryor and Collins (1992).

seen where, and also which way the animal is facing – a type of interpretation which in some cases involves perceiving the animal as a perceiver, and marking a distinction between which parts of the scene it can perceive and which it cannot.

An important fact about the duck-rabbit picture is that people report that it *looks* different in the two views. This is not the same as seeing one thing, treating it as evidence, and then drawing two different sorts of mutually exclusive conclusions, like a detective finding evidence that narrows the set of suspects down to two individuals. Neither is it like looking at an object, and then being reminded of two other objects. The phenomenology of the duck-rabbit ambiguity, along with many other familiar ambiguities, for instance the old-woman young-woman picture Figure 6, shows that *how things look* as opposed to *how things are thought about*, or *what they remind us of* can change even when there is no geometric change, as in the duck-rabbit, or in addition to geometric changes, as in the old-woman young-woman picture, where there are 3-D shape changes as well as changes of more abstract kinds.



Figure 6: *Young-old woman*.

The more abstract, non-geometrical changes involve the identification of parts of a body or face, which way the individual is facing, what the individual can see, and other attributes of an animal or human, such as age, beauty, state of mind, and so on. Exactly what it means to say that these things are aspects of how things look is not clear, but we can, as suggested in Sloman (1989) partially explain this as follows. Saying that something is seen as opposed to merely being inferred implies that the information obtained has a spatial structure, and that spatial parts of the perceived structure are “in registration with” parts of the two dimensional visual field whose structure is determined by the optic array. This has implications such as that shifts of attention between parts or aspects of what is perceived use mechanisms that are related to mechanisms involved in attending to different spatial parts of a complex object.

Of course, in the more complex and subtle cases where what we see is an abstract mental state, e.g. happiness, sadness, elation, dejection, there is no way for that abstract state or its component features to be “in registration” with anything spatial. In this case the location of the state we see is itself something more subtle, as these pictures show.

## 5.2 Seeing or inferring?

Saying that something is seen rather than inferred assumes that there is a distinction between perceptual mechanisms and central cognitive mechanisms, even if they overlap and even if each functions with the aid of the other. Some people may wish to argue that there is no distinction, for instance that the brain is a single unified dynamical system. I shall leave it to neuroscientists

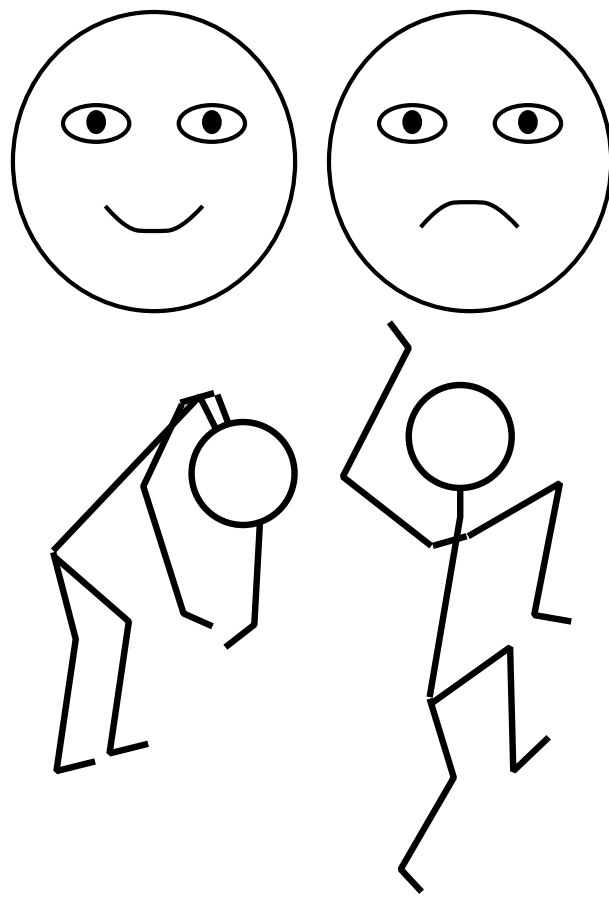


Figure 7: *We can see sadness or happiness in a face, in a posture, in movement. How can one information-processing system see another as happy or sad? What does it mean to say that it does? What does it MEAN to say that affective states are SEEN – not just INFERRED from or associated with the image?*

and others to refute that view, and merely suggest here that insofar as there is a visual sub-system capable of providing information about geometrical and physical aspects of the environment and insofar as the visual mechanisms can make use of learnt associations, for instance between 2-D features of the optic array and 3-D features of objects in the environment, the same mechanisms may be capable of being trained to provide information of other, more abstract kinds, possibly communicated to different central subsystems. This was one of the main claims of the 1989 paper, subsequently partly confirmed empirically by Goodale and Milner (1992).

Thus there is a fairly clear empirical hypothesis that some non-physical, non-geometrical, information is produced by visual *perceptual* mechanisms, as opposed to merely being inferred from results of perception or triggered in a general-purpose associative memory by results of perception. Much of the research done by psychologists on vision and many of the examples presented in psychology text books of visual ambiguities are consistent with this hypothesis.

In a system which already has a distinction between dedicated perceptual mechanisms and general-purpose reasoning mechanisms it makes sense to ask whether a specific task is performed in one or the other. If the perceptual and reasoning mechanisms overlap, it could still be the case that information derived from visually perceived features and relationships is stored in special-purpose information structures that relate the details of the abstract interpretations to specific locations in

a detailed representation of the 2-D structure of the original image or optic array as, happens with other more clearly *visual* contents, as illustrated in Figure 2.

In other words the explanation of the phenomenology of *seeing* the two interpretations of an ambiguous figure is that mechanisms specific to visual processing produce both geometric and non-geometric interpretations and that the more abstract, non-geometric details, including information about affordances (and the ‘reference features’ discussed in Pryor and Collins (1992)) are stored, like the 3-D geometric details, *in registration with* the image features and relationships that give rise to them. Exactly how this is done remains an open question.

If this were being implemented in a computer model we could imagine various ways in which parts of a complex data-structure representing the analysis of the image structure and parts of a complex data-structure representing the inferred 3-D and functional entities and relationships could have large numbers of mutual pointers. (This was done around 1976 in an implementation by David Owen and myself of the “Popeye” system which produced interpretations of dot pictures of the sort shown in Figure 2). These structures and the corresponding pointers can be rapidly computed and discarded as visual attention moves around a complex scene. It is not at all clear how the corresponding functionality can be achieved in brains: neural connections cannot be rapidly discarded and re-grown, for instance, so the theory implies that the brain somehow implements a virtual machine able to support rapidly changing information structures representing spatial structures. (See Trehub (1991) for some suggestions as to how this might be done.)

Of course, it’s one thing to make suggestions regarding how to represent the information and the information relationships. It’s quite another to say how all the information and structures can be computed quickly during a rapid sequence of foveations and how much is preserved between each fixation to help maintain information about the surrounding environment, and how the information is *used* in selecting and guiding actions or in learning generalisations about the environment.

Computer-based systems are still nowhere near being able to combine all these capabilities as far as I know. Moreover, many researchers in vision assume that the task of explaining and modelling human vision is far simpler than I have suggested, because they explicitly or implicitly accept either Marr’s position (described below) or a view that vision can be done entirely through recognising patterns in images and learning correlations within image sequences.

## 6 Visions of vision

The discussion so far has proposed that vision involves a very wide range of types of functions and can have deep connections with cognitive mechanisms of various kinds. However, a great deal of research in vision has been based on a much narrower and more precise view of vision as a process something like reversing the projection of light from surfaces of objects to retinal images. A particularly clear and influential exponent of this view was David Marr, who died tragically in 1981, but continued to have a powerful influence as a result of posthumous publication of his book Marr (1982).

## 6.1 Seeing according to Marr

Everyone agrees that visual perception, at least in humans, is a process that includes operations on a 2-D image (or sequences of images) such as feature-detection, segmentation, grouping, learning correlations between image features and structures, along with recognition of segmented objects. In more sophisticated cases, it can also include detection and representation of 3-D geometrical structures, relationships and motion, along with surface properties such as colour, texture and illumination. Some forms of grouping, segmentation, recognition and learning may be applied not to the 2-D patterns but to the derived 3-D structures and motion. An early example of such work is reported in Hogg (1983).

It is also commonly assumed that that is all that vision involves. Perhaps the most sophisticated version of this type of viewpoint can be found in Marr (1982), where, on page 36, he describes the function of vision as essentially concerned with information about spatial structure and relationships. It is a “*quintessential fact of human vision – that it tells about shape and space and spatial arrangement.*” He admits that “*it also tells about the illumination and about the reflectances of the surfaces that make the shapes – their brightnesses and colours and visual textures – and about their motion.*” But he regards these things as secondary, since “*... they could be hung off a theory in which the main job of vision was to derive a representation of shape*”.

Viewpoints similar to Marr’s have led to a huge amount of research on vision in AI (and also psychology), but the considerations presented in this paper show that it ignores important biological functions of vision, for organisms with diverse needs and capabilities.

## 6.2 What Marr left out: Gibsonian affordances and mental states of others

Our discussion above shows that even if a machine has a complete grasp of the 3-D structure and motion occurring in a scene it will not yet have a human-like view of the scene if it does not also see various possibilities and constraints (impossibilities): including possibilities for acquiring more information by changing viewpoint or moving objects, and possibilities for achieving goals in the environment by moving things in various ways, and constraints on both of those. In addition, human vision involves categorising states of others, as happy, sad, quizzical, attentive, bored, angry, embarrassed, looking left, looking right, and so on.

We can use Gibson’s word “affordances”, in more general way than Gibson did, as a generic label for these counterfactual items of information about things in the environment, information about what or would not happen under various circumstances.

The affordances should not be thought of as *intrinsic*, i.e. *viewer-independent*, aspects of the environment as physical properties and relations normally are. That is because for a particular organism, the perceptual mechanisms will have evolved to provide information tailored not only to what is in the environment but also to the actions which that type of organism can perform and the needs or goals of the organism. The actions will depend in part on its physical structure. So affordances are *relational* properties of scenes and which affordances exist will depend on what they are being related to.

An animal that can grasp things with its tail or capture prey with a long sticky tongue that can be shot out at moving objects will not necessarily perceive the same collections of affordances as a human. A nest-building bird that uses only mud and feathers to build its nest will not see the same

possibilities for action as one that assembles its nest from rigid twigs and sticks that will not simply remain in place if pressed on the edge of the growing nest, but have to be woven in. A monkey that occupies the same sort of tree as a mother orang-utang will not necessarily see the affordances that the orang-utang uses to make a bed for the night by grasping and bending several branches with one hand and holding them in place with its body while holding an infant with the other hand. Even if the monkey could *in principle* perform the same actions, if it never actually builds such sleeping platforms it will not have goals that give those actions sufficient importance for the perception of their possibility to be part of its visual repertoire. In the case of a trainable animal, like a monkey, the perceivable affordances will change over time as part of its development of new skills.

Humans seem to have a very general and versatile ability to learn to see new affordances relating to a huge variety of different sorts of tasks. Skilled tennis players, boxers, pole-vaulters, polo-players, mountaineers, orienteers, hurdlers, carpenters, roof repairers, violin makers probably all acquire different collections of abilities to see and use affordances. However they probably also share a vast collection of such abilities that come simply from being embedded in the same sort of physical world, having roughly the same sort of body, and interacting from infancy through early childhood with the same collection of materials. There may be minor differences arising out of cultural differences in toys, play-materials, and physical disabilities and deformities.

A particular manifestation of human versatility is the fact that people who have been born with structural abnormalities, e.g. the stunted limbs characteristic of babies whose mothers were given thalidomide during pregnancy, often grow up with a wide range of skills other humans do not have. For instance there are artists without arms who can use a paintbrush held with their toes to make excellent pictures, and others who hold the paintbrush with their teeth.<sup>14</sup> An implication of the theory propounded here is that some of the detailed affordances learnt by such people will not be shared by all humans, and some of the affordances detected by people who have normal limbs will not be perceived by thalidomide victims.

The suggestion is that through interacting often with the environment using particular collections of actions we learn to “read” the possibilities inherent in particular objects or configurations of objects in something like the way learning a foreign language can involve learning to read text written in that language. For someone who cannot read the language the characters will still be perceived as marks on a surface, or even as a form of writing. But they will not be immediately and involuntarily recognised and interpreted, as the phrase “a big red ball” is for a fluent English reader. We are all constantly “reading” things in the environment that we have found useful to detect instantly.

The fact that different people learn to perceive different affordances related to their cultural backgrounds, their needs and goals, and their physical capabilities does not prevent them from integrating their integrate their specific collections of perceptual and other capabilities within a common general framework that enables them to live happily within a society where others do not all perceive the same set of affordances,

For animals that can manipulate objects mainly using their mouth or beak (e.g. birds, and most grazing and hunting animals) there is an important fact that the viewpoint will be constantly changing as the grasping and moving occurs, whereas for monkeys, apes and humans the head and eyes can remain still while hands manipulate objects. Often this means that a wider range of spatial and structural relationships can remain in view during the manipulation. This difference may lead

---

<sup>14</sup>See <http://www.indiabuildnet.com/mfpa/index.php>

to important differences in the visual mechanisms and the types of affordances seen and used.

### 6.3 Perceiving empty spaces

It is common to describe an open doorway, or a gap between two large objects as an example of a specific kind of affordance: *passage*. This illustrates a general point mentioned previously: there are significant types of perception where what is seen involves absence of something physical rather than presence: e.g. lack of any obstruction, so that the possibility of motion is perceived.

A less often noticed fact is that a blank sheet of paper, or canvas can be seen as a space in which many things are possible. Which possibilities are perceived will depend on the perceiver. A young child who can barely draw will not see the same possibilities as an experienced adult artist. There is a more abstract collection of possibilities presented to someone who uses the surface not for drawing pictures but for composing music, writing letters, solving algebraic problems or writing computer programs.

An interesting question for our time is what happens to children who grow up spending a great deal of time manipulating objects on a computer screen using a mouse. Does that mean that they will see the screen differently from people who have grown up using such things as paper, pencil, crayons, erasers, paint and paint-brushes?

## 7 The “contents” of visual perception

The discussion so far has had a number of themes, one of which is that vision involves the detection of hypothesized entities of varying scales, varying degrees of abstractness, forming parts of ontologies with different properties and relationships, not all of them spatial or physical. Another theme, generalising Gibson’s notion of ‘affordance’ is that vision involves the detection of a variety of possibilities associated with various locations in the perceived scene, including possible changes that can occur in the environment, possible moves that the perceiver may make in order to gain new visual information, and possible actions the perceiver and other active agents may perform in the environment.



Figure 8: *Some people cannot see the mistake, even when they know there is one.*

There is an interpretation of all this according to which all the visual information is *immediately* and *completely* available to the perceiver. That is what it “feels like” to unreflective self-observers. However, there are many familiar facts that indicate that this is a misleading impression. For instance, there are many cases where we simply do not notice something we are looking for, although it is in full view. People proof-reading text often fail to notice errors on the text, and in some cases may fail to see a mistake even if they have been told that there is one, and it is clearly visible. A well known example is in Figure 8. Another well known example is the blind spot: this corresponds to a part of the visual field in which small items cannot be seen. This information gap

is normally not noticed, even if one eye is shut, so that it is not possible for the other eye to provide the missing information. However, simple experiments described in many text books on vision can make people become aware of something like a small cross disappearing from view if it moves into the blind spot.

These and other phenomena, including “change blindness” in contexts where large changes in an image or scene are simply not noticed, have led some philosophers and scientists to conclude that our experience of simultaneously experiencing a broad, completely filled in visual field, with a great deal of detail is some kind of illusion. E.g. see the recent special issue of *Journal of Consciousness Studies* edited by Alvar Noë Noë (2002).

A different view of the situation emerges from our discussion so far. It might be thought that the multi-layered interpretation of dot-pictures of letters forming a word, as depicted in Figure 2 must always give a complete specification of all the details at every level in the hierarchy. However, we have seen that it is not normally possible for a visual system to have all the relevant information about everything in the scene, and we have suggested that in addition to abstract high level interpretations that do not specify the precise low level details an interpretation of visual information can include information about what to do in order to get more information instead of only information about what is in the optic array, and in its interpretation. In some cases, especially where decisions have to be taken very quickly, the ability of the visual system to “jump to conclusions” about the high level interpretation of the scene, leaving details unspecified and unchecked may be crucially important.

This does *not* imply that the unnoticed details are not taken in. They may be retained in some lower level information structure for a short time in case they are needed. For example, I have found that some people who do not see anything wrong with the familiar phrase in Figure 8 even after they have been staring at it for a few minutes and they know that others see something and are amused can be shown to have taken in the information required to see it. For example, if such a person is asked to close his eyes, and then is asked a question like “How many words were there?” or “Where was the ‘THE’?” then, in some cases, merely considering a different question from the one they originally answered by seeing a phrase, makes them notice a feature of what they saw which they previously did not notice. They apparently notice this by inspecting, with their eyes shut, a lower level information structure whose contents were previously not attended to in detail. (This does not work with everyone. Some people have to look at the picture in order to see what they have not noticed. These individual differences are worth exploring, but I shall not speculate about the reasons.)

There are other cases where behavioural evidence shows that information has been taken in even though people are totally unaware of the fact, for instance if artificially induced changes in optical flow make people fall over even though they have no visual experience of a change in optical flow.

All of this implies that for the purposes of a science of psychology it can be a serious mistake to focus attention on the question “What do I see?” as if there were a unitary entity taking in visual information. I have argued elsewhere e.g. Sloman (1989, 2001b, 2002b, 2001a) that we need to construe visual perception as a process involving a rich, multi-layered, perceptual architecture connected with various parts (reactive, deliberative and reflective parts) of a rich central architecture, at least in humans. From this perspective what people say they see, and what they think they see, or what the experience themselves as seeing, may be just a part, even a distorted part, of what is actually going on in the whole system.

To be continued....

## 8 Betty Crow: cognitive agent and hook-maker

Two crows, Betty and Abel, learnt to use bent wire to fish a bucket of food out of the vertical tube (as in the picture). Then Abel flew off with the hook.

Betty tried using a straight piece of wire for a while, and failed. She then pushed one end of the wire into the tape holding the tube and moved the other round with her beak, making a hook, which she used to lift the bucket as shown in Figure 9.

She did this 9 times out of 10, when tested by being given only a straight piece of wire.<sup>a</sup> How can we find out what Betty was doing? To find more, give google: betty crow hook

---

<sup>a</sup>Reported in Nature and shown on BBC TV (August 2002).

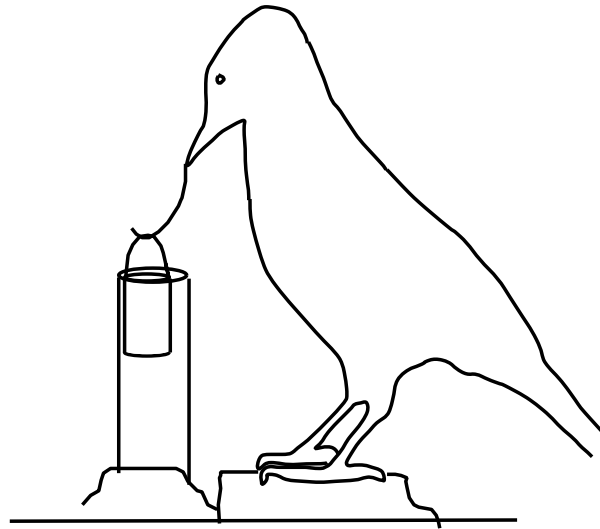


Figure 9: *Betty the hook-making crow.* See the video here: <http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm>

### 8.1 What sort of architecture could do what Betty did?

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

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

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

- The possibility of the bucket moving up the tube,
- The possibility of the wire being bent and holding its shape (unlike most natural objects crows are likely to encounter, which are either elastic and return to something like their original shape after bending, or else brittle and break when bent).
- The possibility of various steps in the process of bending the wire
- The possibility of using the bent wire (which does not yet exist) to lift the bucket of food.

These are all cases of the perception of affordances, i.e. the *possibilities for* and *constraints on* action and change in a situation. In addition, it looks as if Betty either perceived or reasoned about the affordances that did not yet exist but could come to exist as a result of her actions.

## 8.2 What should amaze us?

Most people are amazed when they discover what Betty was able to do. Why? Why are we not equally amazed that a human child (or adult) can see the possibility of a hook-making process inherent in a straight piece of wire, and use that possibility in achieving something that is initially unachievable?

We are not amazed in part because we do not attempt to ask: how could I design a robot that can do such things? One consequence of attempting to design robots that can perform tasks we take for granted in humans is that we become amazed that humans can do them: being amazed requires an appropriate prior education.

Many explanations of Betty's behaviour, with varying degrees of plausibility, are compatible with *any* observed performance, e.g.:

- Pure chance?  
This is highly implausible, and the more often Betty repeats the performance, the less plausible this is.
- An innate behaviour triggered by some mixture of internal and external state?
  - What mixture?
  - How did the genes get the information? Why was this information selected? Was there something analogous to bendable but inelastic wire in the evolutionary history of crows?
- A learnt adaptation in a trainable (altricial) reactive system?
  - What sort of boot-strapping could achieve this?
  - What sort of innate mechanism was required to make it happen?
  - How is the learnt information acquired, represented, stored, activated, used - including information about the properties of pieces of wire?
  - Is it possible that Betty is unique because somehow a piece of wire with which she could play was made available to her during her previous development?
- Was it a deliberative (e.g. problem-solving) process?
  - Using what sort of ontology for possible goals, states, actions?
  - Using what general knowledge?
  - Invoked how?
  - Acquired how? (Using an architecture built in infancy?)
  - Using what planning mechanisms?  
(Using what representations, what search mechanisms?)
- Did it involve self-knowledge? (Reflection/meta-management)
  - Did Betty understand what she was doing, or did she, like many AI deliberative systems, lack reflection/meta-management? (Can a crow teach another crow to do this?)

The questions are deep and important because understanding of spatio-temporal processes can be re-used in many contexts. E.g. doing mathematics, designing architectures, thinking about anything complex.

### 8.3 What a child cannot see

Perhaps we also need to look at more things people cannot see, at first, then later learn to see, and try to understand what changes in between.

An example is shown in this video of a child, Josh, aged about 18 months, playing with a toy wooden train with trucks connected using hooks and rings:

[http://www.cs.bham.ac.uk/~axs/fig/josh34\\_0096.mpg](http://www.cs.bham.ac.uk/~axs/fig/josh34_0096.mpg)

It is clear from the video that although he can see well enough to pick up a truck, put it down, turn it round, move the two metal parts together, he simply does not understand that when he brings the two rings together he is doing the wrong thing. There is independent evidence that he can poke an object into a hole, though that is not shown in the video. It is difficult to tell exactly what is going on, but it looks as if he can see that there is no way to join the trucks together, but he does not understand what he has to do to remedy the situation, so he merely gets frustrated, expresses his frustration, and tries to do something else.

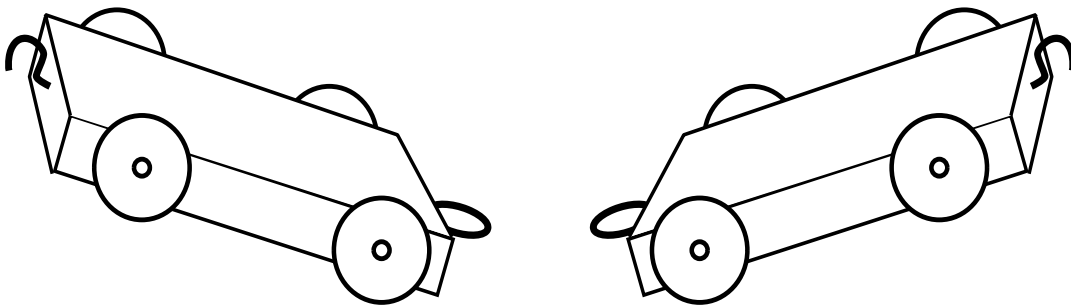


Figure 10: *How would you have to move the trucks to join them together? Why?*

A few weeks later he was able to deal competently with the hooks and rings, although nobody had explicitly taught him (according to his parents). What changed? Of course it depends whether he had blindly somehow learnt a sequence of movements that solved the problem or whether he had somehow *understood* the affordances provided by the different shapes in different positions and relationships. Eventually, a normal human child will understand.

My question is not

- What proportion of children of varying ages, sexes, cultures, social backgrounds, etc. can or cannot solve the problem?
- How long does it take for a child to learn how to solve it?
- What external factors trigger or hinder understanding?
- Is there understanding or blind performance of a learnt sequence of actions?

Rather, I am asking

- What changes within the child between not understanding and understanding? Are there new mechanisms, new forms of representation, a revised ontology with new concepts, new factual knowledge using old or new concepts, new manipulative skill, or perhaps a mixture of all of these?

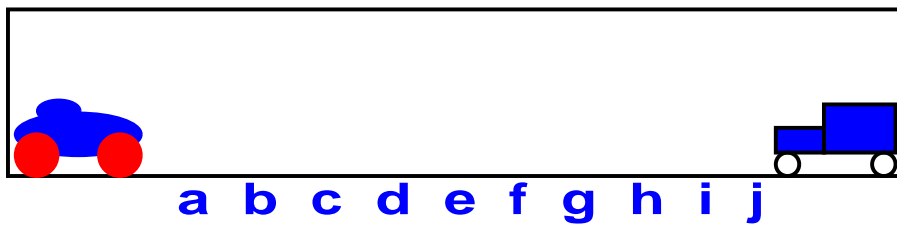


Figure 11: *The two vehicles start moving towards one another at the same time. The racing car on the left moves much faster than the truck on the right. Whereabouts will they meet?*

A different example, requiring understanding of the relationship between space and time is a question I asked a five year old.

One five year old, confronted with the problem posed in Figure 11, using two toy vehicles on a window-sill answered by pointing to a location near ‘b’.

Me: WHY?

Child: IT’S GOING FASTER SO IT WILL GET THERE SOONER.

Again the question I wish to ask is not an *external* question about the age at which a child makes such a mistake, what proportion of children of various sorts makes such mistakes, what external conditions enable them to acquire the ability to produce the right answer, whether the learning can be speeded up or hindered by any educational process, etc. Rather my question is:

*What changes internally between the time when he produces the above reasoning and the time when he understands why that is wrong, and what internal mechanisms make such a change possible?*

Is this another example where the ability to see space as a medium in which changes can occur over time play a role? Can acquiring the ability to visualise two things moving at different speeds in opposite directions play a role. Or is it a problem to be answered only by abstract logical/algebraic reasoning: at any time the distance travelled by the faster vehicle will be greater. Therefore at the time they meet the faster vehicle will have travelled further. Therefore they will meet nearer the truck’s starting point.

## 8.4 Vision and affordances

Vision is often thought to include image and scene segmentation, object recognition, perception of geometrical and physical structure and motion building cognitive maps for small scale or large scale route-planning. It includes various combinations of these functions for different animals, and for humans at different stages of development.

However, there’s something else, which is a deeper, more mysterious capability, evidenced by Betty the crow, usually unwittingly taken for granted in humans, but not yet properly characterised. This is what we have been referring to as “perception of 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)

Affordances in a complex scene can, as suggested above, be construed as

(1) *sets of sets* of counterfactual conditionals,

(2) *spatially indexed*: different sets are associated with different parts of objects.

But this still leaves open what sorts of mechanisms, architectural configurations of mechanisms and forms of representation can explain the ability to perceive them. In particular it is not clear how animal brains represent counterfactual possibilities. Do they use something like modal logics Steedman (2002), or is there some powerful new form of representation waiting to be discovered? Could they be built out of condition-action rules?

## 9 Humans can think with spatial structures.

One consequence of our ability to see possibilities for change in physical structures is that we can use this in reasoning, by manipulating physical structures just as we manipulate sequences of sentences in reasoning logically Sloman (1971).

### 9.1 An example

How do you try to answer the question in Figure 12?

In Figure 12 there are no points common to the triangle and the circle. Suppose the circle and triangle change their size and shape and move around in the surface. They could come into contact. Clearly if a vertex touches the circle, or one side becomes a tangent to the circle, there will be one point common to both figures. If one vertex moves inside the circle and the remainder of the triangle is outside the circle how many points are common to the circle and the triangle? What are all the possible numbers of common points? How do humans answer the question?

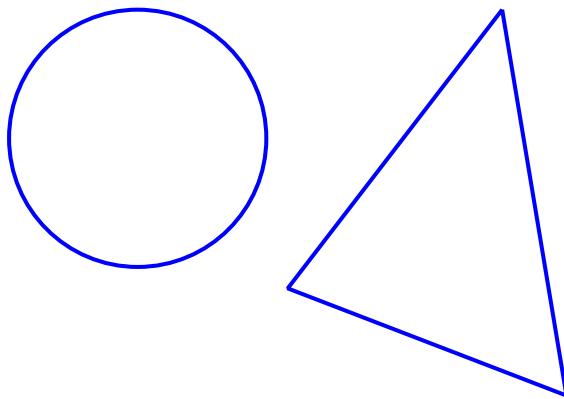


Figure 12: *How many contact points can there be?*

### 9.2 Can we believe introspections?

People often claim on the basis of introspection, sometimes supported by (admittedly controversial) laboratory experiments (e.g. by Kosslyn and many others) to be able to manipulate images of

some kind when solving problems. Sometimes it is claimed that the experience is exactly like the experience of seeing a physical image. These claims should not necessarily be taken literally.

For instance a person who claims to be able to visualise familiar words as if they were written on a wall may be shown to be able to read the letters in a familiar word (e.g. “grandfather”) backwards far more quickly if the letters are actually on the wall than if they are merely visualised as being on the wall.

This may suggest that visualised structure far from being implemented in a kind of flat isotropic 2-D display may actually be implemented in something like a linked list structure, where each item is in a link directly associated with the successor link but not the predecessor link. (See chapter 7 in Sloman (1978)<sup>15</sup>)

(ADD discussion: Reasoning with infinite ordinals.)

The main basis of my criticism of logicians in 1971 was the oft-noted fact that human beings fruitfully use many different forms of representation (Hayes 1974 calls them ‘schemes’), including natural languages, gestures, maps, musical notation, dance notation, Venn diagrams, Euler diagrams (often confused with Venn diagrams!), dress-making patterns, programming languages, blueprints, flow-charts, 3-D models of molecules, many types of data-structures used in computing systems, and a host of special-purpose mathematical notations (including, for instance, the number notation in which concatenation of digits stands for the combination of multiplication by 10 and addition).

More recently, connectionist representations have also added to the variety. For instance, a neural net may include information about many instances of some category superimposed and distributed over a collection of synaptic weights.

expressibility vs heuristic power vs economy of expression e.g.

Confusions about efficiency

analogical vs diagrammatic/spatial

My 1971 paper defined analogical representations as those that used properties and relationships in the representing medium, rather than explicit symbols, to represent properties and relationships in the situation depicted. This was misinterpreted by some readers as the claim that analogical representations are always isomorphic with what they represent, even though the paper presented a counter-example in the form of a 2-D picture of a 3-D scene.

all sorts of other types, including those mentioned above. In Sloman 1993a, 1993b, 1994b, 1994c, I have begun to develop the notion of a mind as a self-modifying, self-monitoring, control system, and to generalise the concept of ‘representation’ to cover a host of different types of information-bearing, causally effective, control states. These control states can exist in virtual machines at different levels of abstraction, like function definitions in a Lisp virtual machine or rules in a Prolog virtual machine or OPS-5 virtual machine. From this general viewpoint, all sorts of different types of representations, internal and external, can be seen as having syntax, semantics, pragmatics and inference methods. Syntax is a matter of the available structures and forms of variation. Semantics is concerned with one thing representing (depicting or denoting) another. Pragmatics is concerned with the functional roles of various sub-systems in a larger whole. Inference methods are simply the pragmatically useful syntactic transformations.

---

<sup>15</sup>Also available here <http://www.cs.bham.ac.uk/research/cogaff/crp/chap8.html>

I believe that attempts to understand and replicate human-like *visual* capabilities will not succeed without some radically new forms of representation that integrate information about spatial structure and motion with information about possible changes, causal relations, and functional roles in a deep way (Sloman 1989). The fact that we do not yet know how to give machines visual capabilities that even begin to match the sophistication of human vision, or even squirrel vision, is one reason why it has been hard for research in AI to make use of the obvious fact that visual representations play a powerful role in human (and animal) intelligence.

we sometimes use what we *see* as a basis for reasoning or problem solving, e.g. looking at the shape of an armchair and its relationship to a door, in order to work out how to get it through the door by rotating it. In such cases we do not choose the representation: the environment presents us with it, and all the hard problems of vision are there.

## 10 How can we increase HMS?

The

Human vs computer (program):

cognitive powers (what does that mean?):

languages (e.g. varieties of syntactic forms understood)

ontologies, reasoning abilities,

abilities to learn through doing and interacting,

visual and other perceptual abilities, goals,

ability to ask the other questions or explain things to the other

physical ability to access

Computer cannot see the screen, usually,

Human cannot see internal data-structures, usually

physical ability to change the contents of the interface, ....

(Some differences more interesting than others. Some harder to remove than others.)

Limit the display to include only elements and structures that both the computer and the user can change

Limit the display to include only elements and structures that both the computer and the user can read

Limit the display content to something both human and machine can understand (limit semantic content to something both can understand) includes restriction of ontologies

Limit goals and tasks to those either could in principle aim for and achieve

Increase cognitive powers of computers (long term goals of AI)

# 11 Simple examples

Programmable Editors: Vedit and Emacs - considerable symmetry at the textual level extendable by user programs: unlike most editors

[POSTPONE OR REORGANISE OR REMOVE NEXT TWO SECTIONS]

## 11.1 Varieties of limitations of current machine image understanding

Sometimes the computer does not even grasp the same syntactic structure as we do. For instance, many software packages generate diagrams or images, or even synthetic videos, which the user can see and understand but the machine cannot. In some cases it will have no ability to decompose the images in the way that a human does, for instance into parts of a human body, or different objects in a cluttered scene. If the computer generates a picture of two rotating wire frame cubes, one behind the other, it may not detect the changing intersection points of the lines depicting the edges of the cube.

Even if a TV camera is placed in front of the screen and the input is fed back to a program in the computer which does detect moving lines and their intersections, it need not notice that some intersections in Figure 5 correspond to corners of cubes where edges meet and others correspond to “accidental” crossing points due to two line segments in space projecting to the same plane surface. Understanding the difference requires a grasp of the notion of depth, the difference between 2-D and 3-D structures, and the notion of a “line of sight”.

The lack of visual competence can seriously limit the usefulness of the machine in many applications. For instance, there are many online libraries of images where it is impossible to ask the computer to find an image satisfying a description (e.g. “one containing at least three people dressed up for a happy occasion”, or “one showing a horse-drawn vehicle”) unless humans have previously studied the images and provided annotations that can be used by search processes.

If a machine can be trained to do such searches by being shown labelled examples of different sorts of images from which it constructs a neural net or some kind of statistical analyser, if trained using current techniques it will typically not understand the import of the human descriptions of the images matched. E.g. it may match images on low level 2-D features whereas what we are matching are 3-D interpretations which happen to be reliably correlated with those low level features within a certain range of scenes, though not all. Partial occlusion, for instance, can upset the regularities. For a statistically trained recognizer that will just be a kind of noise. For a viewer who understands 3-D structures and lines of sight the missing portions in the image will be understood and expected.

## 11.2 What is it to understand the scene?

If the computer is displaying an exploded view of some complex machinery, the user may point at something and ask “What is this for?” and get no sensible answer because the computer cannot tell what is being pointed at, even if there is a well-defined contact point, between finger and screen.

A request to rotate the object so as to make a part more visible to the user may not be understood, for instance because the computer does not know what makes something visible to a person, and therefore cannot rotate something using a grasp of spatial structures and lines of sight to tell when

the relevant part is visible from a particular direction. Of course a programmer may anticipate such questions and put specific code into the package to deal with them, but the machine will not thereby given an understanding of what it is doing which could be used to answer questions like “How should the object be rotated so as to make that part visible to someone standing on the left?”

When the computer does not understand, a user may have to learn ways to move a mouse or twiddle knobs or press “arrow” keys to get the object in the image rotated appropriately. The machine will be able to interpret these actions as instructions to alter some numbers or other information in data-structures which will have the side-effect of changing what is on the screen. But the machine will not normally understand the purpose of the changes nor the effect they have on what the user can or cannot see or understand or do.

Very young children form an intermediate case: they can see and, to some extent, understand images of three dimensional objects but without necessarily having an older child’s ability to understand what is visible from the viewpoint of another person nor what difference that can make to another person’s ability to think and act, for instance the ability to work out how to move a hand to grasp something, and then to do it.

There are many other questions that current computers would be prevented from answering and tasks they would be prevented from performing because of their limited understanding of scenes. A person can tell whether the duck or the rabbit in Figure 5 can see the cube, or where the cube would have to be moved in order to be visible to the rabbit. This requires an understanding of what X can see, where X is an intelligent agent in the scene, and perhaps knowledge of the typical field of view of members of particular species: a hunting bird, such as an owl, has more forward facing eyes than a duck. This is related to which way X is seen to be facing, but can be complicated by the presence of obstacles or mirrors.

Other examples of scene understanding involve perception of causal or functional relationships, e.g. whether X is supporting Y, whether a moving object is likely to collide with another object, whether one rotating cogwheel will make another rotate clockwise or anti-clockwise, what will happen if a string is cut, and so on, which way one should move in order to catch a ball, and so on. (In Sloman (1971) it was suggested that geometric relations between parts of a scene could be used to search for causal connections relevant to answering such questions.)

Much of the information involved in perceiving a causal or functional relationship is inherently concerned with counterfactual conditionals. But most work on how to represent scenes does not address the issue of how to represent counterfactual conditionals.

## **Acknowledgements**

Some of this work was done during a project funded the Leverhulme Trust on ‘Evolvable virtual information processing architectures for human-like minds.’ I have benefited from conversations with many colleagues including Jane Riddoch, Glyn Humphreys, Ron Chrisley, Manuela Viezzer, and Jeremy Wyatt.

# References

- Bar-Hillel, Y. (1964). A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation. In *Language and Information: Selected Essays on their Theory and Application*, pages 174–179. Addison-Wesley, Reading, Massachusetts.
- Barrow, H. and Tenenbaum, J. (1978). Recovering intrinsic scene characteristics from images. In Hanson, A. and Riseman, E., editors, *Computer Vision Systems*. Academic Press, New York.
- Beaudoin, L. (1994). *Goal processing in autonomous agents*. PhD thesis, School of Computer Science, The University of Birmingham. (Available at <http://www.cs.bham.ac.uk/research/cogaff/>).
- Beer, R. (1998). Framing the debate between computational and dynamical approaches to cognitive science (commentary on ‘The dynamical hypothesis in cognitive science’ by T. van Gelder). *Behavioral and Brain Sciences*, 21(5):630.
- Dennett, D. (1987). *The Intentional Stance*. MIT Press, Cambridge, MA.
- Gibson, J. (1986). *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, Hillsdale, NJ. (originally published in 1979).
- Goodale, M. and Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25.
- Hogg, D. (1983). Model-based vision: A Program to see a walking person. *Image and Vision Computing*, 1(1):5–20.
- Marr, D. (1982). *Vision*. Freeman.
- Minsky, M., Singh, P., and Sloman, A. (2004). The St. Thomas common sense symposium: designing architectures for human-level intelligence. *AI Magazine*, 25(2):113–124. <http://web.media.mit.edu/~push/StThomas-AIMag.pdf>.
- Minsky, M. L. (1978). A framework for representing knowledge. In Winston, P. H., editor, *The psychology of computer vision*, pages 211–277. McGraw-Hill, New York.
- Minsky, M. L. (1987). *The Society of Mind*. William Heinemann Ltd., London.
- Minsky, M. L. (1992). Future of AI Technology. *Toshiba Review*, 47(7). <http://web.media.mit.edu/minsky/papers/CausalDiversity.html>.
- Neisser, U. (1967). *Cognitive Psychology*. Appleton-Century-Crofts, New York.
- Noë, A. (2002). Is the visual world a grand illusion? *Journal of Consciousness Studies*, 9(5–6):1–2. (Editor’s introduction to special issue).
- O’Regan, J. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24:939–1031.
- Pryor, L. and Collins, G. (1992). Reference features as guides to reasoning about opportunities. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, pages 230–235, Bloomington, Lawrence Erlbaum Associates.

- Scheutz, M. and Schermerhorn, P. (2002). Steps towards a systematic investigation of possible evolutionary trajectories from reactive to deliberative control systems. In Standish, R., editor, *Proceedings of the 8th Conference of Artificial Life*. MIT Press.
- Searle, J. (1980). Minds brains and programs. *The Behavioral and Brain Sciences*, 3(3). (With commentaries and reply by Searle).
- Sloman, A. (1971). Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd IJCAI*, London. Reprinted in *Artificial Intelligence*, vol 2, 3-4, pp 209-225, 1971, and in J.M. Nicholas, ed. *Images, Perception, and Knowledge*. Dordrecht-Holland: Reidel. 1977.
- Sloman, A. (1978). *The Computer Revolution in Philosophy*. Harvester Press (and Humanities Press), Hassocks, Sussex. Online at <http://www.cs.bham.ac.uk/research/cogaff/crp>.
- Sloman, A. (1985). What enables a machine to understand? In *Proc 9th IJCAI*, pages 995–1001, Los Angeles.
- Sloman, A. (1987). Reference without causal links. In du Boulay, J., D.Hogg, and L.Steels, editors, *Advances in Artificial Intelligence - II*, pages 369–381. North Holland, Dordrecht.
- 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 at <http://www.cs.bham.ac.uk/research/cogaff/>).
- Sloman, A. (1990). Notes on consciousness.
- Sloman, A. (1992). The emperor's real mind. *Artificial Intelligence*, 56:355–396. Review of Roger Penrose's *The Emperor's new Mind: Concerning Computers Minds and the Laws of Physics*.
- Sloman, A. (1993). The mind as a control system. In Hookway, C. and Peterson, D., editors, *Philosophy and the Cognitive Sciences*, pages 69–110. Cambridge University Press, Cambridge, UK.
- Sloman, A. (1994). Explorations in design space. In Cohn, A., editor, *Proceedings 11th European Conference on AI, Amsterdam, August 1994*, pages 578–582, Chichester. John Wiley.
- Sloman, A. (1996a). Actual possibilities. In Aiello, L. and Shapiro, S., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)*, pages 627–638, Boston, MA. Morgan Kaufmann Publishers.
- Sloman, A. (1996b). Towards a general theory of representations. In D.M.Peterson, editor, *Forms of representation: an interdisciplinary theme for cognitive science*, pages 118–140. Intellect Books, Exeter, U.K.
- Sloman, A. (1997). What sort of control system is able to have a personality. In Trappl, R. and Petta, P., editors, *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*, pages 166–208. Springer (Lecture Notes in AI), Berlin.

- Sloman, A. (2000a). Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?). In Dautenhahn, K., editor, *Human Cognition And Social Agent Technology*, Advances in Consciousness Research, pages 163–195. John Benjamins, Amsterdam.
- Sloman, A. (2000b). Interacting trajectories in design space and niche space: A philosopher speculates about evolution. In M.Schoenauer, *et al.*, editor, *Parallel Problem Solving from Nature – PPSN VI*, Lecture Notes in Computer Science, No 1917, pages 3–16, Berlin. Springer-Verlag.
- Sloman, A. (2001a). Beyond shallow models of emotion. *Cognitive Processing: International Quarterly of Cognitive Science*, 2(1):177–198.
- Sloman, A. (2001b). Evolvable biologically plausible visual architectures. In Cootes, T. and Taylor, C., editors, *Proceedings of British Machine Vision Conference*, pages 313–322, Manchester. BMVA.
- Sloman, A. (2002a). Architecture-based conceptions of mind. In *In the Scope of Logic, Methodology, and Philosophy of Science (Vol II)*, pages 403–427, Dordrecht. Kluwer. (Synthese Library Vol. 316).
- Sloman, A. (2002b). How many separately evolved emotional beasts live within us? In Trappl, R., Petta, P., and Payr, S., editors, *Emotions in Humans and Artifacts*, pages 35–114. MIT Press, Cambridge, MA.
- Sloman, A. and Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, 10(4-5):113–172.
- Sloman, A., Chrisley, R., and Scheutz, M. (2004). The architectural basis of affective states and processes. In Arbib, M. and Fellous, J.-M., editors, *Who Needs Emotions?: The Brain Meets the Machine*. Oxford University Press, Oxford, New York. Online at <http://www.cs.bham.ac.uk/research/cogaff/sloman-chrisley-scheutz-emotions.pdf>.
- Sloman, A. and Chrisley, R. L. (2004). More things than are dreamt of in your biology: Information-processing in biologically-inspired robots. *Cognitive Systems Research*.
- Sloman, A. and Croucher, M. (1981). Why robots will have emotions. In *Proc 7th Int. Joint Conference on AI*, pages 197–202, Vancouver.
- Sloman, A. and Logan, B. (2000). Evolvable architectures for human-like minds. In Hatano, G., Okada, N., and Tanabe, H., editors, *Affective Minds*, pages 169–181. Elsevier, Amsterdam.
- Sloman, A. and Scheutz, M. (2001). Tutorial on philosophical foundations: Some key questions. In *Proceedings IJCAI-01*, pages 1–133, Menlo Park, California. AAAI. <http://www.cs.bham.ac.uk/~axs/ijcai01>.
- Steedman, M. (2002). Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, 25:723–753.
- Treuhub, A. (1991). *The Cognitive Brain*. MIT Press, Cambridge, MA.

Wright, I., Sloman, A., and Beaudoin, L. (1996). Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101–126. Repr. in R.L.Chrisley (Ed.), *Artificial Intelligence: Critical Concepts in Cognitive Science*, Vol IV, Routledge, London, 2000.

Zadeh, L. A. (2001). A New Direction in AI: Toward a Computational Theory of Perceptions. *AI Magazine*, 22(1):73–84.