Architectural and Representational Requirements for Seeing Processes, Proto-affordances and Affordances.

Aaron Sloman

University of Birmingham, School of Computer Science,
Edgbaston, Birmingham B15 2TT, UK
http://www.cs.bham.ac.uk/~axs/
A.Sloman@cs.bham.ac.uk

May 30, 2008

Abstract. This paper, combining the standpoints of philosophy and Artificial Intelligence with theoretical psychology, summarises several decades of investigation by the author of the variety of functions of vision in humans and other animals, pointing out that biological evolution has solved many more problems than are normally noticed. For example, the biological functions of human and animal vision are closely related to the ability of humans to do mathematics, including discovering and proving theorems in geometry, topology and arithmetic. Many of the phenomena discovered by psychologists and neuroscientists require sophisticated controlled laboratory settings and specialised measuring equipment, whereas the functions of vision reported here mostly require only careful attention to a wide range of everyday competences that easily go unnoticed. Currently available computer models and neural theories are very far from explaining those functions, so progress in explaining how vision works is more in need of new proposals for explanatory mechanisms than new laboratory data. Systematically formulating the requirements for such mechanisms is not easy. If we start by analysing familiar competences, that can suggest new experiments to clarify precise forms of these competences, how they develop within individuals, which other species have them, and how performance varies according to conditions. This will help to constrain requirements for models purporting to explain how the competences work. For example, Gibson's theory of affordances needs a number of extensions, including allowing affordances to be composed in several ways from lower level proto-affordances. The paper ends with speculations regarding the need for new kinds of information-processing machinery to account for the phenomena.

Keywords. Vision, affordances, architectures, development, design space and niche space, dynamical-systems, evolution, functions, generalised-languages, mathematics, proto-affordances, representations, scaling up *vs* scaling out,

Dagstuhl Seminar Proceedings 08091
Logic and Probability for Scene Interpretation
http://drops.dagstuhl.de/opus/volltexte/2008/1656

This paper is available at:
http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0801a
Slides available here:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk59

A much compressed version of this is to be included in *Proceedings of Computational Modelling Workshop, Closing the Gap Between Neurophysiology and Behaviour: A Computational Modelling Approach* Editor: Dietmar Heinke
http://comp-psych.bham.ac.uk/workshop.htm
University of Birmingham, UK, May 31st-June 2nd 2007

1 From Kant to Gibson and Beyond

1.1 What are the Functions of Vision?

The purpose of the workshop for which this paper was prepared was to discuss a computational approach to "Closing the gap between behaviour and neurophysiological level". My approach to this topic is to focus almost entirely on what needs to be explained rather than to present any neurophysiological model, though conjectures regarding some of the design features required in such a model are offered in Section 7.

This is one of a series of interim reports on a journey towards understanding human vision as forming a subset of a large collection of competences within an integrated multi-functional, self-extending information-processing architecture: a whole mind. Over several decades I have been trying, as a philosopher-designer, to understand what the functions of vision are, and how vision relates to the rest of the information-processing architecture, especially in humans but also in other animals. This paper adds some observations arising in part from work on a project to explore requirements and possible designs for a robot that can perceive and manipulate 3-D objects. The main outcome has been expanding the list of human visual capabilities that we still do not know how to explain.

My initial motivation came from trying to understand the role of visual processing in mathematical discovery and reasoning, for instance in proving theorems in elementary Euclidean geometry, but also in more abstract reasoning, for example about infinite structures, which can be visualised but cannot occur in the environment. This work started in my DPhil [1], which was an attempt to defend the view of mathematical knowledge as both non-empirical and synthetic, proposed by [2], but rejected by many contemporary mathematicians and philosophers. This is a topic that links many disciplines, including mathematics, psychology, neuroscience, philosophy, linguistics, education and, since the 1950s, Artificial Intelligence.[1]

[1] For details see [3] and "Could a child robot grow up to be a mathematician and philosopher?" (PDF presentation) available online here:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#math-robot

1.2 Vision's Role in Mathematical Discovery

I shall try to show how the role of vision in mathematical reasoning is connected with the ability in humans and some other animals to perceive and reason about structures and processes in the environment, including *possible* processes that are not actually occurring.

Studying vision's role as the basis for important human mathematical competences focuses attention on aspects of vision that are ignored by most other researchers including: those who study vision as concerned with image structure (e.g. [4]); those who study vision as a source of geometrical and physical facts about the environment (e.g. [5]); those who regard vision as primarily a means of controlling behaviour (e.g. [6] and many recent researchers in AI/Robotics and psychology who regard cognition as closely tied to embodiment); and those who regard vision as acquisition of information about affordances (e.g. [7,8]).

The work of James Gibson on affordances comes close to identifying competences related to mathematics, but I shall show that his theory has to be broadened in various ways, especially since perception of affordances depends on perception of what I call proto-affordances (Sections 4, 4.2 and 5.1), which include actual and possible processes not necessarily produced by the perceiver. The ability to represent combinations of proto-affordances is important for perceiving more complex affordances, for predicting and planning future processes, for explaining past events and for understanding *vicarious affordances* (Section 4.3). The ability to perceive and reason about combinations of and interactions between proto-affordances (Section 5.8) is the core of the connection between functions of vision and mathematical competences, as explained in later sections.

Analysis of examples reveals further details, including the need to be able to use an "exosomatic" ontology referring to things in the environment (e.g. 3-D surface and processes involving them) as opposed to patterns in sensory and motor signals. Other competences requiring extended ontologies are also mentioned, e.g. the meta-semantic ability to see mental states of others, such as emotional states (Section 6.6).

For a full understanding of what evolution has achieved and what future robots may have to do, we need parallel investigations of many different kinds of vision: in insects and other invertebrates, in birds, in primates, etc. There are some who believe that all research should start from the simplest systems and work up, but it is better to avoid both dogmatism and narrowness, and see what can be learnt by explorations pursing different directions, as long as the explorers communicate.

1.3 Scientific Communication Problems

Communication requires shared concepts, however, and that can be a problem. Much philosophical research is connected with the fact that we can have a collection of words that we are very familiar with, and use successfully in day to day communication, but whose mode of operation is far more complex than we realise, because they correspond to concepts whose structure is not obvious (e.g.

"truth", "knowledge", "understanding", "consciousness", "justification", "science" and "emotion"). As a result, when such words are used to define scientific objectives or report results of scientific investigations, researchers often fail to notice disparities between their everyday use of the concepts and their professional use: disparities that can obstruct the advance of knowledge, by obscuring the differences between what researchers have actually achieved, and their claims to have modelled phenomena they describe as "seeing", "learning", "understanding language", "feeling emotions", or "being conscious". A similar point was made by [9], in a criticism of many AI research reports.

I shall illustrate this by showing that our ordinary concept of "seeing" covers everyday human achievements that go far beyond the topics normally studied in empirical and computational research on vision. In particular, seeing is not restricted to information about what exists in the environment: Seeing what is possible or impossible (discussed in more detail in [10], and below in sections 4 and 6) is different from seeing what exists, and requires different mechanisms and forms of representation. J.J. Gibson's (1979) notion of perception of affordances turns out to be a special case.

1.4 Why Complete Architectures Matter

Ordinary phenomena of seeing inherently involve interactions between sensory mechanisms, action-control mechanisms and more central systems that arise from different stages in our evolutionary history and grow during different stages in individual development. So the functions of vision differ from one species to another and can change over time within an individual as the information-processing architecture grows. Some of those developments, such as what language the individual learns to read, or which gestures are understood, are culture-specific.

A full understanding of vision requires investigation of different multifunctional architectures in which visual systems with different collections of competences can exist.

An architecture with more sophisticated 'central' mechanisms can make possible more sophisticated visual functions. For instance, a central mechanism able to use an ontology of causal and functional roles is required for a system that can see something causing, preventing, or enabling something else to occur. The ability to make use of an ontology including mental states (a meta-semantic ontology) is required if a visual system is to be able to perceive facial expressions, such as happiness, sadness, surprise, etc. and make use of the information.

The requirement to use such ontologies is ignored by machine vision researchers who train pattern recognisers to attach labels to pictures on the basis of 2-D image features. Linking labels such as "happy" and "sad" with image features, without any understanding of the causes of happy or sad mental states or their likely consequences, does not constitute seeing someone as looking happy or sad. Likewise, seeing 3-D structures and processes, and causal interactions between them, requires the use of an ontology that refers to contents of a 3-D environment, which is quite different from being trained to use a set of labels for 2-D images or image sequences generated by such an environment. Understanding a 3-D ontology involves being able to ask and answer questions about surfaces and volumes, about spatial relations, about kinds of motion, about kinds of interaction, and being able use the information in planning or controlling actions, and being able to explain perceived events. More subtly, it can involve discovering the need to extend the ontology beyond what can be sensed, e.g. to include different physical substances and their properties. These requirements are discussed further in Section 5.

1.5 Links to Philosophy

Many philosophers who study problems relating to human minds and human knowledge, focus mainly on trying to clarify how our current concepts work, for example [11] (who described conceptual analysis as studying "logical geography" of concepts), and work referred to in [12]. There is an alternative approach. Instead of simply analysing how our ordinary concepts referring to mental phenomena work, since learning about AI and the design-based approach to understanding minds, around 1969, I have been trying to characterise deeper aspects of the functions performed by human minds (including their visual sub-systems). This exposes a more varied collection of functions than our ordinary concepts categorise and distinguish. [13] describes the underlying space of possibilities as defining a "logical topography", on the basis of which we can show that familiar sets of concepts provide only one among many "logical geographies", just as a portion of terrain can be divided into political or social regions in different ways.

By analysing the different sets of functions supported by different information-processing architectures we can come up with a theory-based survey of possibilities for a mind or a visual system. For each system design there is a specific "logical topography", a set of possible states, processes and causal interactions that can occur within the architecture, some involving also interactions with the environment.

The set of possibilities generated by each architecture can be subdivided and categorised in different ways for different purposes. E.g. the purposes of common sense classification are different from the purposes of scientific explanation.

If we adopt the design-based approach when observing actual performances by adult humans, infants, toddlers, nest-building birds, squirrels, and other animals, and constantly ask "how could that work", we can generate various collections of requirements for information-processing architectures and mechanisms that could support the observed variety of visual functions in robots. Very often, it is not obvious whether a particular theory meets that criterion: so using a theory as a basis for designing, implementing and testing working artificial systems is a crucial part of the process of explaining how natural systems work.

This sort of enquiry can often reveal serious confusions and over-simplifications in our ordinary concepts, even though they work well enough most of the time in non-scientific contexts. For example, notions like "learning", and "memory" are normally used in ignorance of the variety of ways in which complex information-processing systems can acquire and use information, and adapt themselves to changing circumstances, on various time scales.

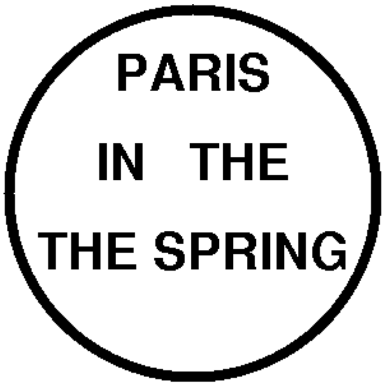

Fig. 1. *Many people see nothing wrong with the contents of this circle, but can be shown to have acquired all the information, without making use of it.*

1.6 The Logical Geography of Our Talk About Seeing

Similar comments can be made about the ordinary use of "see". For example, many people think it is impossible to see something without being conscious that you have seen it, yet there are cases where people perform actions that require seeing things even though they would not be described as being conscious of what they see. Opening a door while sleep-walking, and not remembering stopping at a red traffic light while driving to work are examples.

Figure 1 gives another example. Some people cannot see anything wrong with the phrase displayed even when pressed repeatedly to look very carefully. A subset of those individuals can then be asked while their eyes are shut "Where was the 'the'?", or "How many words were there?". At that point some of them notice the mistake in the phrase. Presumably some part of them did see what was there, and stored the information but did not use it until it was required for answering a different question about the contents of the display. Similar results are available from many laboratory experiments in visual and auditory perception (e.g. dichotic listening experiments). We can clarify different concepts of seeing by analysing architectures that do, and architectures that do not support what could be called "unconscious seeing".

Another example is the common assumption that, when not hallucinating, we see only things that exist in the environment. That may be a correct assumption as regards the functions of vision in insects and other animals. Yet, as will be explained in Sections 4, 4.2 and 5.1, seeing affordances, and seeing the lower-level proto-affordances they include, involves seeing the possibility of actions that can be performed and processes that can occur, and also seeing constraints on such possibilities. This involves seeing that some things that do not exist in the environment can or cannot occur in that environment. If an architecture supports that kind of seeing, it is misguided to look for retinal projections of what is seen: something that does not exist cannot be projected to form an image.

Gibson's work can be seen as making this sort of point, though I shall try to show that his objections to naive notions of the functions of vision do not go far enough, because he did not see that affordances are a subset of a more general class (implicitly acknowledged after his death in [8]).

There are many other conceptual confusions related to notions like "feeling", "consciousness", "emotion", "understanding", and "attention" that can be clarified in terms of different sets of capabilities supported by different architectures [14]. Researchers working on functions of mind and vision often focus on a narrow range of functions, without asking how the functions studied fit into a complete architecture.

Many specific visual functions have been modelled in working systems, but currently available AI techniques still leave large gaps, and all existing models will probably be viewed as toys when we look back at them fifty or a hundred years from now, because they account for such a small subset of human or animal competences. That criticism applies also to my own system-building experiments, done with colleagues and students (including the Popeye system mentioned below).

Although designing and testing working systems is often informative, what may be of longer-lasting value, as suggested in [15], is assembling ever-expanding sets of *requirements* and sketching out a sequence of very high level designs for meeting more and more of those requirements, while testing the designs wherever possible both by model building and by deriving predictions that can be checked in laboratories or in fieldwork. However, outline designs for systems meeting a large collection of requirements are very difficult to implement at present.

This paper should be viewed as work-in-progress, reporting some of the results of extended exploration of requirements for human-like visual systems. It describes some unobvious, yet important, functions of vision that need to be explained by specifications of a working machine that could serve those functions in addition to the more obvious functions. The paper ends with some speculations about sorts of mechanisms (using a large collections of multi-stable, interlinked dynamical systems) that may be required, and which do not yet exist in any known computational models, and which may be hard to identify in neural mechanisms, using current observational techniques.

2 The Need to Compare Alternative Designs

2.1 Niche Space and Design Space

Understanding how a complex system works includes knowing what would happen if various aspects of the design were different, or missing. So understanding how humans work requires us to relate the human case to many others, namely other products of biological evolution and possible future engineering products. The comparison with different biological designs extends a familiar theme for neuropsychologists, namely attempting to understand normal human functions by comparing them with various effects of brain damage or genetic brain abnormality.

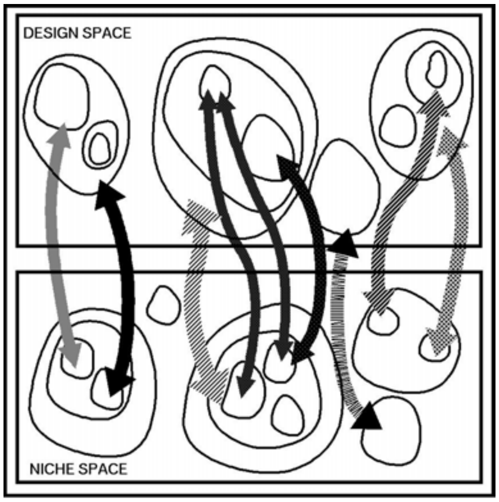

Fig. 2. *To understand human capabilities we need to know what requirements and constraints they satisfy, i.e. where they fit in niche space, and also what sorts of designs for information-processing systems can meet those requirements, i.e. where we fit in design space, and how those spaces are related.*

We can summarise this as follows (building on [16,17,18,19,20]): We need to study the space of possible sets of requirements or niches, *niche space*, and we need to study the space of possible designs for working systems that can meet different sets of requirements, *design space*. And as suggested in Figure 2 we need to understand the various relationships between regions of design space and regions of niche space. Of course, in their full generality these two spaces are far too large to be studied, so we must find ways of homing in on appropriate "neighbourhoods" in those spaces. Unfortunately, it is often tempting to do that in a way that is strongly influenced by the mechanisms and formalisms we are familiar with, which can cause us to be blind to some of the things that need to be explained. An example discussed later is the failure of researchers to notice that much of what we see consists of *processes* in the environment, because our current tools are much better suited to investigations of perception and recognition of *static structures*. Another common hindrance is that many researchers know only how to build software that manipulates numerical information, so they ignore the requirement to build visual systems that are capable of producing *structural descriptions* of perceived entities, processes and unrealised possibilities (including affordances). Moreover, the currently available computational tools for creating and manipulating complex information structures do not seem to be up to some of the tasks described below in connection with spatial reasoning. By examining requirements that are often ignored we may accelerate development of suitable information-processing mechanisms.

Our notion of niche space is not the same as the notion of *niche space of a type of organism* used by biologists, a notion that is concerned with the space of possible environments a particular kind of organism can live and reproduce in. Our notion covers sets of requirements for all possible organisms and robots.

The study of niche space and design space should help us understand how the two spaces are related: which regions of design space map onto which regions of niche space, and in which ways different designs meet or fail to meet particular sets of requirements, or meet them more or less well. This requires a *descriptive* notion of biological fitness, which specifies how the competences provided by an implemented design relate to the various competences specified in a set of requirements.

This is much richer and more complex than any *numerical* notion of biological fitness measured in terms of survival value, or number of progeny. No numerical measure that produces a linear ordering of cases is adequate for understanding any family of complex designs: the numerical information, or position in a ranking, loses far too much detailed information about specific benefits and disadvantages of various design features in different niches to be a useful evaluation criterion – like most numerical evaluation functions used in computational models of evolution or learning. When a design involves cooperating parts, assigning a number, or ranking to the whole design does not provide information about the strengths and weaknesses, or even the functions, of the component designs. Contrast the production of numerical values with the detailed information about advantages and disadvantages of various alternative products in consumer research reports, e.g. in *Which?* magazine.

Full understanding of particular subsets of design space and niche space requires us to explain the pressures that lead to changes over time as systems in those subsets both evolve across generations and support development and learning within individuals. So we also need to understand *trajectories* in both spaces, some of which are evolutionary trajectories of species, some developmental trajectories of individuals, and some social or cultural trajectories. The evolutionary trajectories can include changes within some components of an organism for which other components define the niche. Some evolutionary developments primarily involve changes in behaviours of organisms rather than physical structures, though physical changes may follow. A broader view would also take in trajectories followed by ecosystems containing many species.

2.2 Varieties of Representation: Generalised Languages (GLs)

Complex systems may differ in many ways, some of which are described later. A particularly important feature of any information-processing system is how it is capable of encoding information during various stages of acquisition, analysis, storage, retrieval, and use of the information. Researchers designing computational models often have commitments to particular forms of representation,[2] since those are the ones for which they have programming tools. Those commitments can severely restrict the kinds of research questions they ask, and the answers they consider.

For example, many researchers who use neural nets will make heavy use of vectors of numerical values, matrices for transforming vectors, and algorithms designed for controlling such numerical transformations. A different kind of researcher will be more inclined to use symbolic structures such as trees and graphs, for example trees used to represent syntactic structures of sentences or plans, and graphs used to represent maps, partially ordered plans, or the structures in images.

There are many scientific disputes regarding whether non-human animals can learn to use a language. However these disputes are posed in terms of a notion of language that has certain features common to human languages, including (a) structural variability of sentences (sentences can have more or less complex syntactic structures, with different levels of nesting), (b) compositional semantics (the meaning of a complex whole depends systematically on the meanings of the parts and how they are assembled, which allows novel meanings to be expressed or understood), (c) use of linear sequences of arbitrary symbols to form sentences and (d) the use of sentences for communication between individuals.

If we drop condition (c), namely use of linear sequences of arbitrary symbols, then we can allow use of spatial structures combined spatially in different ways (as in maps, diagrams, pictures, flow-charts etc.). Dropping that constraint allows us to consider a wider variety of forms of representation that satisfy the first two conditions, namely structural variability and compositional semantics. If we drop condition (d), namely use of symbols for communication, then we can still allow other uses such as thinking, reasoning, planning, or perceiving complex scenes. A further generalisation is to allow semantics of complex wholes to depend not only on constituents and structure but also context. This relaxation is normal for indexicals (linguistic components such as “now”, “that”, “you”, etc). We can generalise the role of context in resolving ambiguity in ways that would take too long to explain here.[3]

The resulting notion of a language, a Generalised Language, (G-language or GL) requiring only conditions (a) and (b) was proposed in [21] and elaborated in [22]. A neural model that excludes the possibility of GLs used internally for various purposes such as perception of processes and structures, planning future

[2] I use the word “representation” to refer to whatever is used to encode information. It could be some physical structure or process, or a structure or process in a virtual machine. It could be transient or enduring. It may be used for a single function or for many functions. See also http://www.cs.bham.ac.uk/research/projects/cogaff/misc/whats-information.html

[3] A partial account is in http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0605

actions, thinking about what might happen, and reasoning about possible consequences of processes in the environment would be incapable of meeting some of the requirements described below. Future robots are likely to need internal GLs of various sorts, possibly including some geometrical and topological components in addition to discrete symbols. Spatial forms of representation have often been proposed as having advantages in certain contexts compared with more logical, or sentential, forms of representation, e.g. [23,24,25,26].

Preverbal children and many non-human animals can perceive and react to processes as they occur. That requires mechanisms providing the ability to represent changes while they happen. Perhaps the same mechanisms, or closely related mechanisms, can be used to reason about processes that are not happening. If some other primates and very young children use internal GLs, that suggests strongly that GLs supporting structural variability and compositional semantics evolved before external human languages used for communication, and that GLs also precede the learning of communicative language in individual humans.[4]

We shall later give several examples of human visual competences, including geometric reasoning competences, that seem to require use of GLs, for example in Sections 4.2, 5.2, 5.8, 6, 6.11, 6.12, and 7. The suggestion that GLs are used for all these purposes, including the representation of processes at different levels of abstraction, poses deep questions for brain science, as we'll see later.

3 Wholes and Parts: Beyond "Scaling Up"

Most of the rest of this paper addresses only a small subset of the problems, concerned with requirements for visual systems. Some high level features of possible mechanisms for satisfying those requirements will also be discussed near the end, in Section 7.

3.1 Putting the Pieces Together: "Scaling Out"

This investigation is motivated by interest in a larger set of problems, not just explaining vision. The larger set of problems includes understanding information-processing requirements for a human-like (or chimp-like, or crow-like) organism to perceive, act and learn in the environment. We also aim to produce design requirements for human-like or animal-like robots, and if possible to sketch some features of designs capable of meeting those requirements.

This means that the designs for mechanisms providing particular competences (e.g. visual competences) must "scale out": instances of a design must be capable of interacting with other components in a larger design that satisfies requirements for the *complete* animal or robot.

This contrasts with the frequently mentioned need to "scale up", namely coping successfully with larger and more complex inputs. Many human competences

[4] As explained in more detail in http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang

do not scale up, including parsing, planning, and problem-solving competences. It is possible to produce a highly efficient implementation of some competence that scales up very well but does not scale out: it cannot be integrated with other human competences. Numerical competences of current computers are an obvious example. Less obviously, the most impressive programs that perform board games do not scale out. Even less obviously, I suspect that none of the current implementations of linguistic processing, vision, planning, or problem-solving can scale out. In particular, they are generally designed to work on their own in some test situation, not to work with one another. More detailed examples of the scaling-out requirement are provided later, e.g. in Sections 5 and 6.4.

3.2 Understanding Tradeoffs: Regions in Design Space

Most AI researchers (and most funding agencies interested in AI) are motivated mainly by the goal of producing new useful machines, and therefore have little interest in this comparative investigation of different sets of naturally occurring requirements and designs, produced by evolutionary and developmental processes in biology. If something works well for a practical application, the fact that it does not scale out, i.e. cannot easily be integrated within a multifunctional architecture combining many human abilities, may be irrelevant to the goals of the designers.

However, ignoring these existence proofs can lead to unsuccessful, blinkered searches for solutions to hard engineering problems.

Like some of the founders of AI, I am more concerned with trying to advance understanding of how humans and other animals work than with building new machines, but unlike most of them I believe this requires trying to understand human information-processing mechanisms, not just in their own terms, but as a special case of something more general that takes different forms in different animals, and in different sorts of possible robots. Thus the goal is not what some people describe as achieving "human-level AI", but something more general.[5]

3.3 Biological *Mechanisms* vs Biological *Wholes*

For several decades the relevance of biological *mechanisms* for AI (e.g. neural nets and evolutionary computations) has been recognised. What is relatively

[5] This cross-disciplinary, cross-species, approach has been promoted in a number of tutorials and workshops, including the Tutorial on "Representation and learning in robots and animals" at IJCAI'05 http://www.cs.bham.ac.uk/research/projects/cosy/conferences/edinburgh-05.html, the AISB'06 Symposium http://www.cs.bham.ac.uk/research/cogaff/gc/aisb06/, the Workshop on "Natural and Artificial Cognition" in June 2007 http://tecolote.isi.edu/~wkerr/wonac/,
this BBSRC workshop http://comp-psych.bham.ac.uk/workshop.htm
and the CoSy project's "Meeting of Minds" workshop in September 2007 http://www.cs.bham.ac.uk/research/projects/cosy/conferences/mofm-paris-07/, among many others.

new in the computational modelling community is moving beyond studying what *biological mechanisms* can do, to finding out what *whole animals*, can do and how they do it, which is one way of studying *requirements* to be met by new designs.

Finding out how information flows around brain circuits connected to the eyes, which has received a lot of attention prompted by new non-invasive machinery for measuring what goes on in brains, does not tell us what the information is, how it is represented, or what the information is used to achieve, which can include things as diverse as fine-grained motor control, triggering saccades, aesthetic enjoyment, recognition of terrain, finding out how something works, or control of intermediate processes that perform abstract internal tasks.

It is possible to make progress in the case of simple organisms where neurophysiology corresponds closely to information-processing functions. Measuring brain processes may be informative in connection with evolutionarily older, simpler, neural circuits (e.g. some reflexes), but many of the newer functions are extremely abstract, and probably only very indirectly related to specific neural events. In those cases, results of brain imaging may show some correlations without explaining what is being done or how it is being done. For example, if speakers of two very different languages with different phonetic structures, different grammars, and different vocabularies are given some piece of information (e.g. that a hungry lion is nearby) the brain processes involved in acquiring that information will be very different, and may differ even for speakers of the same language, depending on what those speakers have learnt and in what order.

Suppose that when two such individuals hear and understand reports with the same semantic content, similar parts of their brains show increased activity: that will not answer any of the deep questions about how understanding works. E.g. we shall be no nearer knowing how to produce a working model with the same functionality.

The journey towards full understanding still has a long way to go, and may even be endless, since human minds, other animal minds and robot minds can vary indefinitely. This multidisciplinary research programme is very different from doing experimental psychology, studying brain mechanisms, or building intelligent machines, yet combines aspects of all of them, along with studies of philosophy, evolution, ethology and linguistics, as illustrated in [21,27,28,22].

A problem for such a programme is the difficulty of evaluating intermediate results. Over-emphasising the need for falsifiable hypotheses can slow down scientific creativity. Instead, as proposed by [29] (who extended some of Popper's ideas), we need to allow that distinguishing degenerative and progressive research programmes can take years, or decades.

3.4 Models That "Scale Out"

It should now be clear that being biologically inspired need not bring any advances. For example, some computational modellers try to derive requirements from results of laboratory experiments on humans. If different highly constrained laboratory tasks produce different reaction times, or if certain changes in a visual task increase error rates, or if performance changes in a certain way as a result of practice, then AI researchers sometimes set themselves the goal of designing working systems that perform the same tasks, while mimicking the time differences, error rate changes and other features of the experimental data. But designing an AI model to match observed performance in such a laboratory experiment leaves open the question: does that model "scale out", i.e. can it be extended to form part of a larger model meeting a much wider range of requirements? Not all mechanisms that perform like part of a system are useful parts of something that performs like the whole system.

If not all the requirements for a machine have been specified, then it may be possible to produce a working design that meets the partial requirements but which cannot be extended to a design that satisfies all the requirements, because it meets the partial requirements in the wrong way.

For example, if you want a machine to model a good human chess player, then part of the requirement is that the machine should be able to win games of chess against good human players without taking much longer than humans do to make each move. We already have such chess playing machines. But if you add other requirements, such as that the player should be able to give advice to a weaker player, not by specifying moves for particular board positions but by playing in such a way as to help the weaker player learn both from mistakes and from successes, then it turns out that the obvious designs that do well as competent chess-players are not easily extendable to meet the further requirements, because they blindly execute algorithms (programmed or produced by training) without knowing what they do or why. Such reflective knowledge is not necessary for winning, if the available machines are fast enough to produce results by exhaustive search. However knowledge about higher level features of various games is necessary for explaining how to win, or how to avoid different ways of losing, and for choosing a style of play that is tailored to a learner's needs.

The ability of a design with functionality F1 to scale out to include new functionality F2 is partly a matter of degree: it depends on how much extra mechanism is needed to provide F2. The more essential use the extra mechanism makes of the original mechanism, the better the original mechanism scales out.

A computer vision system can be very good at being trained to recognise certain classes of objects in images, and to match experimental observations, without being extendable so that it can understand why the same object looks different from different viewpoints, and how the appearance changes as the viewpoint changes. (Note that understanding why could simply involve being able to predict such changes, and being able to plan appearance changes in order to gain new information. It need *not* involve being able to *explain* why, or even to formulate the generalisations in any communication.) Further, insofar as recognition merely involves being able to apply a label to a portion of an image, it need not be extendable so as to allow the machine to see what actions can be performed on different objects or what the consequences of those actions will be.

3.5 Vision and Mathematical Reasoning

Most people would not see human mathematical abilities as relevant to the functions of vision, whereas my interest in understanding vision started when I was doing a DPhil in philosophy [1] attempting to defend Kant's philosophy of mathematics against the then prevalent Humean empiricist thesis that all knowledge that is not empirical must be essentially trivial, like the 'definitional' knowledge that all triangles have three angles and that all prime numbers have no proper divisors. These are analogous to 'All bachelors are unmarried', conveying nothing more than you already know if you understand the words. Many philosophers call such propositions "analytic" [30].

In his *Critique of Pure Reason*, Kant had claimed, in opposition to Hume, that there are ways of discovering new truths that extend our knowledge (i.e. they are "synthetic", not analytic) and which are not empirical. He included truths of arithmetic and geometry, such as the truth that seven plus five equals twelve and the truth that the space occupied by a left hand cannot be superimposed on the space occupied by a right hand, no matter how much they are translated and rotated in 3-D space.

My previous experience as a mathematics student convinced me that Hume and contemporary analytic philosophers were wrong and Kant was right. Discovering or understanding a proof is both different from empirically investigating how the world works and different from reflecting on and rephrasing definitions.

Building on the work of Frege and others, [31] tried to show that although mathematics was non-trivial it could all be reduced to logic (Frege thought that was true of arithmetic but not geometry), whereas I, like many mathematicians, knew from first hand experience that doing mathematics often used spatial reasoning rather than logical reasoning alone.

When trying to prove a theorem, mathematicians frequently use the ability to *see* both structural relationships and also the possibility of and the consequences of changing such relationships. For instance, if you look at a triangle with vertices labelled A, B and C, you can *see* that it is possible to draw a straight line from any vertex, e.g. A, to the point which I'll refer to as M, the middle of the line BC, the opposite side. Even without drawing the line AM, you can see that doing so will produce two triangles sharing a side. You can also see that those two triangles have co-linear sides of the same length: BM and CM, both of which are the same perpendicular distance from the opposite vertex A. From this you can infer that the two triangles must also have the same area, showing that the line AM divides the original triangle into two triangles of the same area, even though in general they will have different shapes.

I am not claiming that every human being can see these things. Clearly very young children cannot. Moreover, the ability may develop only in certain environments.

What exactly 'see' means here requires explanation: some would prefer to say that they 'infer visually', or something like that. The point is that, however we describe the ability, it is connected with visual competences and needs to be explained by any adequate theory of how human vision works. Some readers will already have noticed a connection with Gibson's theory of affordances, discussed in later sections. It turns out that his theory has to be extended to account for the role of vision in reasoning.

Looking at a physical triangle is not necessary. Even when thinking about an *imaginary* triangle rather than one drawn on paper, many people can visualise drawing new lines and can reason visually about the consequences. You probably did that when reading my description of what could be done to the triangle, since I deliberately did not provide a figure. Some mathematicians reading my example will be able to translate the theorem into a logical form, and then work out a derivation from some logical formalisation of Euclidean geometry, such as Hilbert's axiomatisation, but they are rare. I am not denying that it is possible to do geometry only within a logical framework, but it is certainly unusual. Most mathematicians first discover proofs geometrically even if they belong to the mathematical sub-culture that feels duty bound later on to produce a purely logical version. That obligation was challenged in Appendix II of [1], which argued that an extra-logical justification is required for accepting the logical axioms and rules as adequate to the purpose.[6]

Not all mathematical discoveries are based on visual reasoning. For example, very different discoveries, some of them documented in Chapter 8 of [32], occur as a child learns to count, and then (sometimes unconsciously) discovers different uses for the counting process and different features of the counting process, such as the fact that the result of counting a collection of objects is not altered by rearranging the objects but can be altered by breaking one of the objects into two objects. That kind of mathematical discovery depends on perceiving structures and relationships in procedures that can be followed.

Sometimes abstract structures relating the application of procedures, have to be noticed, rather than spatial structures and relationships. For example, a child who has learnt to count may discover that in order to work out the size of the set formed by combining a collection of M objects and a collection of N objects all it has to do is recite N numerals after M. E.g. reciting three numerals after "five" gives "six, seven, eight". A child may discover this for a few cases and then notice that there is a general pattern that can always be relied on. This depends on an information-processing architecture that includes *self-observation* mechanisms that (a) detect features of the processes generated when procedures are applied, (b) work out a common pattern and then (c) notice new instances of that pattern. The ability to understand why the pattern can be relied on always to work requires additional capabilities in the architecture, and usually develops later, as part of the process of becoming a young mathematician. (This process is often terminated by going to school.)

In contrast, simultaneous perception of spatial and temporal relationships can lead to the discovery that any one-to-one mapping between elements of two finite sets can be converted into any other such mapping by successively swapping ends of mappings, as illustrated in Figure 3. This sort of discovery requires quite

[6] Hilbert's axiomatisation of Euclidean geometry is conveniently available at: http://www.math.umbc.edu/~campbell/Math306Spr02/Axioms/Hilbert.html

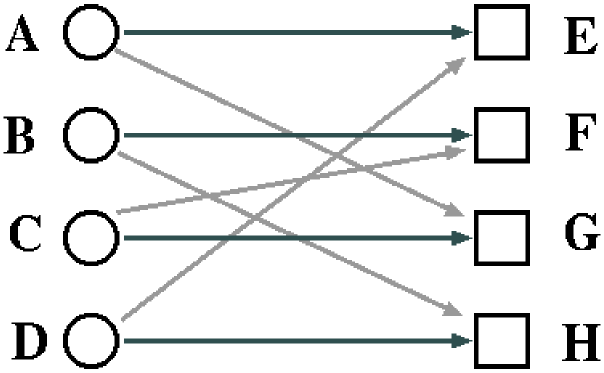

Fig. 3. *A child may first discover empirically that any one-to-one mapping from one set of objects to another (e.g. the grey arrows) can be converted to any other such one-to-one mapping (e.g. the black arrows) by swapping ends on one side, two at a time. E.g. the right hand ends of the grey arrow from A to G and the grey arrow from D to E can be swapped, then the right hand ends of arrows from B to H and from C to F, etc. gradually eliminating discrepant mappings. Formulating the general algorithm is left as an exercise for the reader. How does the visual system find a discrepant mapping? How does it find another to swap with it?*

abstract visual capabilities. For example, vision is needed to detect a remaining discrepant mapping (e.g. the grey link that goes from B to H instead of from B to F, in the figure). Then it is necessary to find a mapping with which to swap it, namely the mapping that goes to F from C. If the right hand ends are swapped the new mapping from B to F can be left thereafter. However the new mapping from C to H will then need to be swapped with one that goes to G.

Initially the procedure might be followed using physical links, e.g. lines drawn on paper that are rubbed out and redrawn, or coloured threads joining objects which can be relocated. Later the young mathematician can *simulate* the process, i.e. imagine doing it, in order to work out the consequences, without actually changing anything in the environment.

A side effect of applying such a strategy repeatedly, seems to be production of implicit understanding that it will always work, even if the child cannot articulate the strategy nor explain why it works.

This seems to depend on the architecture allowing one process to observe that another process has some consequences that do not depend on the particularities of the example, and which are therefore necessary consequences of the procedure. This discovery may use the fact that the visualising or imagining process ignores details that can vary from case to case.

It is worth noting (as argued in [1]) that even using explicit logical reasoning depends on the ability to visualise structural relations between symbolic structures, and possible structural changes, e.g. constructing a new proposition using components of premiss propositions.

3.6 Development of Visual Competences

Many people can see and think about geometrical possibilities and relationships. Very young children cannot see all of them, let alone think about them. Why not? And what has to change in their minds and brains to enable them to see such things? Answering those questions will require explaining how such mathematical visual reasoning works and what needs to develop in order to allow it to work.

There are developments that would not normally be described as mathematical, yet are closely related to mathematical competences. For example, a very young child who can easily insert one plastic cup into another may be able to lift a number of cut-out pictures of objects the child can recognise from recesses, and know which recess each picture belongs to, but be *unable* to get a back into is recess: the picture is placed in roughly the right location and pressed hard, but that is not enough. The child apparently has not yet extended his or her ontology to include boundaries of objects and alignment of boundaries. The problem does not arise for circular cups because of their symmetry. Some time later such a child will easily insert a picture into its recess, presumably after learning about alignment of boundaries.

How such extension of competences happens is not known. Such learning may include at least three related aspects:

- developing new forms of representation;
- extending the learner's ontology to allow new kinds of things that exist;
- developing new ways of manipulating representations for purposes of perception, planning or reasoning, including acquiring new algorithms.

In some cases there is also development of new motor skills making use of the new cognitive competences, e.g. learning to play a musical instrument or play a competitive game.

When all of those developments have occurred and the new extensions have been used a lot, many special cases of their use can be developed and stored for rapid retrieval and use, allowing new problems to be solved far more quickly than before. Such components can then be building blocks used in further developments. Learning to read text or music illustrates all of this very well.

In the case of the child playing with puzzle pieces, what has to be learnt, namely facts about boundaries and how they constrain possible movements, is something that can be studied mathematically. Presumably what the very young child learns is a precursor to being able to think mathematically about bounded regions of a plane. Later mathematical education will build on general abilities to see structures and processes and see how some structures can constrain or facilitate certain processes, as illustrated in [33]. This is related to, but more general than, learning about affordances, as explained in the next section.

4 Affordance-related Visual Competences: Seeing Processes and Possibilities

4.1 Perceiving and Reasoning About Changes

Visual, geometrical, reasoning capabilities depend on several visual competences, such as: (a) the ability to attend to parts and relationships of a complex object, including "abstract" parts like the midpoint of a line and relationships between widely separated objects or features, such as collinearity, similarity of size, or parallelism, (b) the ability to discern the possibility of changing what is in the scene, e.g. adding a new line, moving something to a new location, altering a relationship between two or more objects, (c) the ability to work out the consequences of making those changes, e.g. working out which new structures, relationships and further possibilities for change will come into existence if those changes are made.

One of the consequences of making a change is the production of a new set of possibilities for change and constraints limiting further changes. For example, drawing a new line across an old line produces a new point of intersection, and new bounded regions in the neighbourhood of that point. These new features can then be the basis of further changes, e.g. colouring a region, labelling the point, using the point as the centre of a circle, etc. Being able to see such possibilities in diagrams was part of the geometric reasoning capability discussed in [23], and explored further in [16]. Later it became clear that such mathematical capabilities are closely related to perception of what Gibson called "affordances" [7,34].

It is not always noticed that both the ability to see and make use of affordances and the ability to contemplate and reason about geometric constructions depend on a more primitive and general competence, namely the ability to *see processes* (as opposed to merely seeing structures), and the closely related ability to see *the possibility* of processes that are not actually occurring, and also *constraints* that limit those possibilities. It seems unlikely that all animals that have visual capabilities include these abilities to see that various processes are possible even if they do not occur, or to see that some processes are blocked by features of the environment that could be removed in order to allow the processes to occur. This is one of the topics to be investigated in a study of different regions of niche space and design space in biological systems.

Implicit in our discussion so far is the fact that "multi-strand" relationships can hold between two objects, insofar as not only the whole objects are spatially related (e.g. above, near, north-west of) but also parts of the object. Consequently, when objects move that can produce "multi-strand processes", in which many relationships change simultaneously.

Gibson's perceived affordances were concerned only with opportunities for *actions* that *the perceiver* could perform, including moving towards, avoiding, grasping, lifting, pushing, obstructing, catching, throwing, changing viewpoint, etc. Most of these actions involve physical processes that have much in common with processes in the environment that are not intentional actions, for instance objects blown in the wind, or a rotten branch breaking and falling because of its weight. There are also processes that occur when intentional actions are performed by others.

It is clear that most normal humans are able to perceive what is common between processes that they produce, processes that others produce and processes that are not parts of anyone's intentional actions, e.g. two surfaces coming together or moving apart, an object rotating, one object moving into or moving past another, and so on. Being able to represent what is common to all processes satisfying some abstract conditions independently of how then are perceived and whether the agent initiates them imposes important requirements on the forms of representation. It implies, for example, that representation of perceived processes should not be restricted to incipient motor processes, even if incipient motor processes *sometimes* play a role.

4.2 Proto-affordances and Possible Processes

Not only can we see such processes when they actually occur independently of our own actions, we can also perceive and think about the *possibility* of such a process occurring without having to regard it is an action we can produce, or a process that can affect us. You can see a rock on a steep hillside and think about what would happen if the rock started rolling down, no matter what the cause of the rolling. In short, we can see what could be called *proto-affordances*: namely relations in the environment that enable or constrain processes that are possible in a situation. They could become parts of positive or negative affordances for the perceiver under certain circumstances, but need not be regarded as affordances in order to be seen, thought about, or predicted. The notion of a proto-affordance is essentially the notion a situation in which some process can occur or is prevented from occurring, or constrained in some way. This is different from the notion of a micro-affordance, namely an affordance related to a small sub-action of an action involved in an affordance, or a potentiated action triggered by seeing objects that have affordances [35]. Proto-affordances need not involve actions.

The ability to see that certain processes are possible even when they are not within the power of the perceiver to produce, i.e. the ability to perceive proto-affordances, underlies the ability to perceive what we shall call "vicarious" affordances, namely affordances for others. This requires the ability to represent the possibility of things happening in the environment independently of any of the perceiver's goals being achieved or obstructed, and independently of the agent doing anything in the environment. Thinking about future possible processes extends that to representing processes that may occur in situations that are not currently perceived – for instance thinking about what will happen if it rains, or if the wind blows – without specifying any details or any specific viewpoint. In other words, these abilities use an ontology that is not restricted to sensorimotor contents. They need an "exosomatic ontology", the ability to refer to entities and processes that can exist outside the body, independently of any sensory or motor signals. A pre-verbal child or non-verbal animal that can see and reason about such proto-affordances and vicarious affordances is probably using a spatial GL, as suggested in Section 2.2.

4.3 Vicarious Affordances and Exosomatic Ontologies

When perceived possibilities involve what someone else can or cannot do, we can describe them as involving "vicarious" affordances. Learning to see vicarious affordances can be very important for adults whose children need help while they learn ever more complex (and sometimes dangerous tasks), or for animals that need to anticipate or constrain the behaviours of other animals in fighting with them, attempting to catch and eat them, or attempting to avoid being eaten by them. It need not be the case that perception of such vicarious affordances has to be based on being able to use such affordances in one's own actions. An animal that needs to escape from a flying predator by choosing a appropriate shelter need not itself ever have been a flying hunter.

In many animals, instead of an ability to perceive potential negative affordances for predators and make use of them there is an inherited tendency to react to predators by seeking shelter in appropriate locations, e.g. running into burrows. In those animals the problem has been solved by evolution and a fixed solution adopted. A more sophisticated animal might be able to choose between two shelters by assessing their difficulty for the predator, as humans can do.

In more advanced cases, observation of the capabilities of a predator or enemy can lead to a new design for a shelter that is deliberately built for the purpose, as started happening after aircraft came into use in warfare in the 20th century, and potential targets were either camouflaged or protected in bomb-proof shelters. The use of black-outs during night bombing raids were an example where people accepted conditions that removed their own affordances in order to reduce the positive affordances for the enemy. It is an open research question whether other animals can reason about vicarious affordances, though it seems that both pre-verbal human children and some chimpanzees can perceive and react altruistically to affordances for others, as shown by [36].[7] That research emphasises questions about altruistic motivation, whereas I am drawing attention to the representational and conceptual competences that make it possible to be both helpful and unhelpful to others.

4.4 Evolutionary Significance of Independently Mobile Graspers

A feature of the ability to perceive and use affordances that is not often noted is that there are commonalities between affordances related to doing things with left hand, with right hand, with both hands, with teeth and with tools such as tongs and pliers. Compare Figure 4. In principle it is possible that all the means of grasping are represented in terms of features of the sensorimotor signals involved, but the variety of such patterns is astronomical. Even using only one hand, an object can be grasped in many different ways and because hands can

[7] Videos are available at `http://email.eva.mpg.de/~warneken/video.htm`

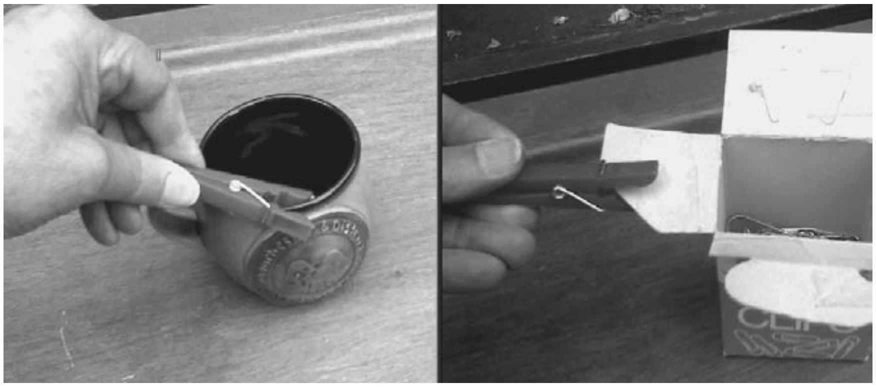

Fig. 4. *Four examples of grasping: two done by fingers, and two by a plastic clip. In all cases, two 3-D surfaces move together, causing something to be held between them, though retinal image projections and sensorimotor patterns in each case are very different.*

move independently of eyes, the variety of retinal projections produced by grasping processes is so great that having to learn all the relevant image structures separately would be a mammoth task.

If, however, grasping is represented more abstractly, in terms of 3-D relations between surfaces in space, using an amodal form of representation, using an exosomatic ontology, i.e. referring to things outside the body instead of only to sensorimotor signals, the variety of cases can be considerably reduced: for instance very many types of grasping involve two surfaces moving together with an object between them until contact is achieved, after which, if the surfaces move together the object moves with them. Sub-cases of that general process-pattern include different object weights and sizes, different kinds of surface (e.g. rigid, compressible, smooth, rough, etc.), flat *vs* curved *vs* articulated grasping surfaces, and so on. The most abstract representation can be used for high level planning of actions involving grasping, including, for example, the common requirement for the two grasping surfaces to be further apart during the approach than the diameter of the thing to be grasped. More detailed information about the specific case can then be used either when planning details, or during action execution to control the detailed motion.

I suspect that biological evolution long ago “discovered” the enormous advantages of amodal, exosomatic, representations and ontologies as compared with representations of patterns in sensorimotor signals. Since the 3-D process-features described in the last paragraph are common to grasping with the left hand, grasping with the right hand, grasping with two hands, grasping with the mouth, and also grasping done by other individuals, enormous economy can be achieved by directly representing the process that occurs in the *environment*, rather than learning and storing myriad different relations between motor signals that produce grasping, and sensor signals that result. The ability of infants to transfer information about affordances from one hand or foot to another, and the economy achieved by using more general forms of representation, is mentioned in [8], though they do not discuss the visual and cognitive mechanisms and detailed forms of representation required.

There are also forms of grasping that are more complex than a process in which two surfaces close in on something between them, for instance, grasping of tools or implements for use in particular ways, such as holding a pen for writing, a screwdriver for turning, a knife for cutting, a fork for prodding and lifting food, a wooden ball for bowling, a pair of scissors for cutting, and a baby for bathing or dressing. Those variants are not discussed in this paper, though the perceptual, representational, and control requirements for each of them would add significantly to the points made here.

Although 2-D image projections are often helpful for controlling the fine details of an action during visual servoing (as noted in [16]), using an exosomatic ontology and representing the 3-D spatial structures and processes, rather than using only somatic sensorimotor signal patterns, can make it possible for an individual to learn about an affordance in one situation and transfer that learning to another where sensor inputs and motor signals are quite different: e.g. discovering the consequences of grasping an object with the right hand then transferring what has been learned to other or observing grasping one by another individual and then attempting to produce a similar spatial process.

The possibility of such transfer depends on a general ability to project 3-D processes to possible sensory signals (e.g. working out what to look for when grasping with the left hand for the first time), and to possible motor signals, e.g. working out how to make the grasping happen in a new way. If there are generic, re-usable, mappings between 3-D processes in the environment and sensor and motor signal patterns, then concepts referring to abstract features of environmental processes, and generalisations expressed using such concepts, will be widely re-usable.

Using such a powerful form of representation makes it unnecessary to rely on magical properties of “mirror neurones”, unless those neurones are simply part of the brain’s mechanism for constructing and using amodal exosomatic representations. Which evolved first will not be discussed here.

Another example may be the ease with which we can *feel* the shape of a surface or the depth of a hole by stroking or poking with a firmly held stick instead of relying on contact with fingers. Our brains have developed ways of mapping different sensorimotor patterns into the same kinds of exosomatic representations.

5 Towards a More General Visual Ontology

5.1 More on Proto-affordances

Both affordances for oneself and affordances for others depend on something more fundamental: namely causal relationships that arise from structural relationships between objects or parts of objects in the environment. In Section 4.2 we labelled these "proto-affordances". An animal or machine that can perceive and represent such relationships in the environment, may be able to make far more predictions, explain far more phenomena, and plan far more solutions to practical problems than one that is restricted to representing only information about sensorimotor patterns, or information about its own present affordances for action.

For example, if a ball is in the space between two vertical surfaces of large rocks, the surfaces may prevent the ball moving horizontally in one direction while allowing it to move horizontally in another direction – parallel to the surfaces. If something is moving, and there is a large object in the path, then the presence of that obstacle can prevent the indefinite continuation of the motion. These possibilities for processes and constraints on possibilities for processes exist and can be perceived independently of whether they are relevant to the goals or actions of any perceiver.

We describe causal relationships between objects or parts of objects that make some processes possible and others impossible as "proto-affordances", since they have the potential, in appropriate contexts, to be the basis for affordances of animals or robots that might have goals or might attempt actions that would be enabled or constrained by these causal relationships.

But a particular proto-affordance, such as the potential of one object to impede the motion of another, can be the basis for a wide variety of action affordances. E.g. it could produce a negative affordance for an agent trying to push the moving object towards some remote location. It could provide a positive affordance for some individual wishing to terminate the motion of a moving object, e.g. a mother wishing to obstruct something rolling down a hill towards her infants.

5.2 The Need for Generative Process Representations

An animal or a machine that can discover such proto-affordances, and has means of representing them that allow manipulation of representations, may have the ability to combine a given set of proto-affordances in different ways in order to generate representations of a huge variety of affordances. An animal that can only learn about and store information about positive and negative affordances that it has already encountered will have a far more constrained understanding of what can and cannot occur in the environment, and will be more limited in its ability to think about entirely new processes and structures before it has ever encountered them. This may be one of the main differences between humans and most other animals. However, some non-human animals seem to show evidence of creative problem solving that uses the ability to combine proto-affordances to form new complex affordances. A particularly famous example was the New Caledonian crow Betty, who seems to have invented several ways of transforming a straight piece of wire into a hook in order to lift a bucket of food from a glass tube [37].

It is not known what forms of representation birds or other animals are capable of using when perceiving or reasoning about 3-D processes and proto-affordances. Anyone who tries making a rigid nest in a tree, by using only one hand and adding only one twig at a time should develop a healthy respect for the intelligence of certain nest-building birds.

Most AI vision researchers, and many psychologists assume that the sole function of visual perception is acquiring information about which objects exist in the environment, and what their properties and relationships are, whereas the facts assembled here suggest that another major function of vision (at least in humans, and probably several other species) is to acquire information from the environment about *what does not exist but could exist*. Moreover, if the ability to reason about consequences of possible processes uses the visual representations involved in perceiving actual processes, then one of the functions of visual mechanisms is to do reasoning, presumably using GLs (Section 2.2). This conforms with the experience of many mathematicians and scientists.

It is not clear how much of this ability to construct, manipulate and use representations of actual and possible processes exists at birth (as the ability to run with the herd exists in some mammals at birth) but that manipulative ability certainly develops, as does the ability to see both what actually exists in the environment and what is possible in any given situation. Examples of such learning in human infants and children are presented in [8], although much of the book focuses on how a child learns about affordances for itself and for others rather than more general learning about properties of the environment.[8]

5.3 Complex Affordances: Combining Process Possibilities

One of the important things learnt by a child (or animal) exploring the environment by acting in it, is that affordances can be *combined* to form more complex affordances. This depends on the fact that actions can be combined to form more complex actions, which in turn depends on the more basic fact about the physical environment that *processes can be combined to form more complex processes.* Reasoning about such complex processes in advance depends on the ability to combine simpler proto-affordances to form more complex ones.

Because processes occur in space and time, and can have spatially and temporally related parts, they can be combined in at least the following ways:

- processes occurring in sequence can form a more complex process;
- two processes can occur at the same time (e.g. two hands moving in opposite directions);
- processes can overlap in time, e.g. the second starting before the first has completed;
- processes can overlap in space, for example a chisel moving forwards into a rotating piece of wood;

[8] The developing ability of a child to see new kinds of things by growing an ontology is discussed in more detail in an online presentation on "Evolution of ontology- extension", at `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0604`

- one process can modify another, e.g. squeezing a rotating wheel can slow down its rotation;
- one process can launch another, e.g. a foot kicking a ball.

The ability to represent a sequence of processes is part of the ability to form plans prior to executing them. It is also part of the ability to predict future events, and to explain past events, but this is just a special case of a more general ability to combine proto-affordances. How is this done?

Both pre-verbal children and animals without human language seem to perceive spatial structures and processes and to be able to do some reasoning about spatial relations and processes, e.g. when they solve a problem for the first time. Examples are reported in [36]. They are unlikely to be reasoning by means of a human language, or a logical formalism.

5.4 Generative Forms of Representation

How do animal brains represent a wide variety of processes? Formal grammars are capable of summarising infinitely varied classes of symbolic structures. A grammar for sentences in a spoken language can generate a potentially infinite variety of acoustic processes. There have also been various attempts to produce systematic ways of generating and representing spatial structures. For example, in the 1960s various researchers experimented with grammars for classes of pictures. Later that was followed by work on classes of 3-D structures. One example was the "geon" theory of Biederman, proposing (implausibly) that humans see all 3-D structures as derivable from a small set of primitive objects that can be deformed and combined in various ways [38]. Marr's (1977) theory of 3-D perception based on generalised cylinders was a variant on this sort of theory. Others have proposed alternative ways of generating classes of 3-D structures. [40], generalising earlier work by Huffman, Clowes and Waltz, report on mechanisms for interpreting a wide variety of 2-D depictions of 3-D polyhedra. There have been many experiments with schemes for generating classes of pictures of plant-like, or animal-like shapes.

The demands on a system for representing spatial *processes* are greater than demands on generative specifications of spatial *structures*. There are far more processes that can occur in any environment than static structures, since each structure can be moved or deformed in many ways (translating, rotating, stretching, compressing, shearing twisting, etc.) and any two structures can move in relation to each other in many different ways. That extra complexity is not expressible just by adding an extra dimension to a vector, which suffices for the move from 3-D to 4-D points. So a system that can represent, see, predict and reason about processes will need to be far more complex than mechanisms for coping with static 3-D structures.

The set of possible processes can be constrained by a context, but even then the remaining set can be very large. There are many familiar human and biological contexts that determine distinct vast and varied collections of possible spatial processes, for example a child's playroom, a kitchen, a group of people at a dinner table, a garden, a motorway, various situations in which birds build nests, scenarios in which lions hunt their prey, a lion eating its prey (a collection of processes that change their character as more of the animal has been opened up and eaten), etc. A theory of how biological vision works must explain what kinds of information about spatial processes particular animals are capable of acquiring, how the information is represented, how it is used, how the ability to acquire and use more kinds of information develops, and so on.

5.5 Generative Schemes for Spatial Processes

There are not many formal generative schemes for classes of spatial *processes*. Newton's laws characterise an infinite variety of possible motions of point masses and larger structures. There have also been various attempts at dance notation. Computer programming languages are powerful representations of processes, and some of them are used to represent processes simulating 3-D physical processes. But the representations have hitherto mostly been required for generating graphical displays, and not for supporting perceptual and cognitive processes in a perceiver of the simulated world.

I don't know whether anyone has attempted to produce a generative scheme for combinations of spatial processes that are likely to occur in the environment inhabited by particular sorts of animals, or by a child in a particular culture. In AI, symbolic planning formalisms have been in use since the 1950s, and these provide a means of representing individual events in terms of preconditions and postconditions. Complex and varied combinations of these event representations allow processes to be represented consisting of sequences of events. The importance of such forms of representation was recognised very early, e.g. by [41]. However, discrete sequences are not enough: the notations used in such planning systems are not suitable for representing the many kinds of spatial process that can occur in our environment in which features and relationships change continuously, and changes occur concurrently at different levels of abstraction. For example such planning formalisms cannot capture the process of wiping a sink clean as perceived when someone actually does it, though the formalism will work for an arbitrarily "chunked" summary of the process, e.g. *pick up cloth; wipe sink; rinse cloth!*

It seems that in order to accommodate the variety of processes humans and other animals can perceive and understand, they will need forms of representation that have the properties we ascribed to spatial GLs in Section 2.2, with the benefits described in [23].

5.6 Varieties of Learning About Processes

In the first few years of life, a typical human child must learn to perceive and to produce many hundreds of different sorts of spatial process, some involving its own body, some involving movements of other humans and pets, and some involving motion of inanimate objects, with various causes. Researchers are often amazed at the speed with which young children extend their vocabulary, after they start talking. I suspect that if anyone ever finds a way to count the rate at which a pre-verbal child extends its ontology and its ability to construct and manipulate new representations in its mental virtual machine, we shall be even more impressed.

This learning process almost certainly requires forms of representation and an ontology that are not given at birth (as happens for the vast majority of animals) but are built up in layers, where the processes in later layers can be both more complex and more abstract than the processes represented in earlier layers, and depend increasingly on the geographic, climatic and cultural influences on the child's environment.

For example, topological processes where contact relationships, containment relationships, alignment relationships go into and out of existence are different from metrical processes where things change continuously. Process representations restricted to changes of shape, size, and geometrical or topological interactions between objects are not as rich as representations that refer to kinds of material and their causal and functional roles, e.g. elasticity, rigidity, impenetrability, stickiness, weight, etc.

Moreover, the very same physical process can include both metrical changes, as something is lowered into or lifted out of a container and discrete topological changes as contact, containment, or overlap relationships between objects and spatial regions or volumes change.

5.7 Process-primitives and Compound Processes

In each context there are different sets of 'primitive' processes and different ways in which processes can be combined to form more complex processes. Some simple examples in a child's environment might include an object simultaneously moving and rotating, where the rotation may be *closely coupled* to the translation, e.g. a ball rolling on a surface, or *independent of* the translation, e.g. a frisbee spinning as it flies. Other closely coupled adjacent processes include: A pair of meshed gear wheels rotating; a string unwinding as an axle turns; a thread being pulled through cloth as a needle is lifted; a pair of laces being tied together by a moving hands and fingers; a bolt simultaneously turning and moving further into a nut or threaded hole; a sleeve sliding on an arm as the arm is stretched; and sauce in a pan moving as a spoon moves round in it.

Many compound processes arise when a person or animal interacts with a physical object. Compound 3-D processes are the basis of an enormous variety of affordances. For example, an object may afford grasping, and lifting, and as a result of that it may afford the possibility of being moved to a new location. The combination of grasping, lifting and moving allows a goal to be achieved by performing a compound action using the three affordances in sequence. The grasping itself can be a complex process made of various successive sub-processes, and some concurrent processes – for example concurrently changing relationships between different parts of the surface of the object and different parts of the grasping hand, as metrical and topological relationships change between:

1. palm and fingers
2. the grasped object
3. the spatial envelope of the hand
4. the spatial envelope of the grasped object or object part.

It is often thought that the only significant result of making use of an affordance is producing a physical change in the environment. But an immediate, more abstract, consequence of a physical change is typically the existence of new positive and negative affordances. The handle on a pan lid may afford lifting the lid, but once the lid is lifted not only is there a new physical situation, there are also new affordances: for pouring the contents of the pan, adding other things to the pan, stirring the contents, seeing the state of the contents, etc. The last example illustrates the fact that an action can alter the *information* available in the environment, which is an epistemic affordance, just as the action of moving closer to an open door (Figure 5) alters epistemic affordances.[9]

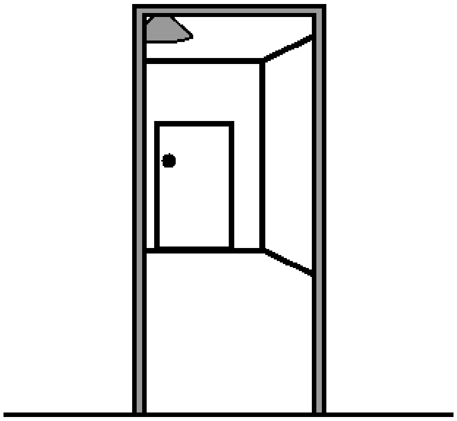

Fig. 5. *As you move nearer the door you will have access to more information about the contents of the room, and as you move further away you will have less. Why? If you move left or right you will change the information available to you. In all these cases, physical actions change the epistemic affordances in a situation.*

5.8 Reasoning About Interacting Spatial Processes

We have seen that processes that occur close together in space and time can interact in a wide variety of ways, depending on the precise spatial and temporal

[9] Further discussion of epistemic affordances and how they change can be found in http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0702

relationships. It is possible to learn *empirically* about the consequences of such interactions by observing them happen, and either collecting statistics to support future predictions, or formulating and testing universal generalisations. However, humans and some other animals sometimes need to be able to consider and *work out* consequences of possible combinations that they have never previously observed, for example approaching a door that is shut, while carrying something in both hands, for the first time. It does not take a genius to work out that an elbow can be used to depress the handle while pushing to open the door.

[8] state on page 180 that the affordance of a tool can be discovered in only two ways, by *exploratory activities* and by *imitation.* There are many unsolved problems about what sorts of mechanism can extract and store re-usable information about how to do something on the basis of observing either other people performing actions or performing them oneself. However the most important point in the present context is that a third way of discovering affordances was not mentioned, namely *working out* what processes are possible when objects are manipulated, and what their consequences will be.

A very strong requirement for human-like visual mechanisms is that they should produce representations (e.g. GLs, if our theory is correct) that can be used for reasoning about novel spatial configurations and novel combinations of processes, which in humans seems often to involve the same kind of reasoning as led to the study of Euclidean geometry long before the development of logic and algebra as we know them. Likewise, young children can reason about spatial processes and their implications long before they can do logic and algebra (as Piaget realised) and to some extent even before they can talk. As remarked in [22], this has implications for the evolution and development of language.

The requirement to be able to use information gained visually to reason about the consequences of novel processes is an important example of the need for designs to scale out, introduced earlier in Section 3.

I assume that by “exploratory activities” Gibson and Pick meant to refer to physical exploration and play, interacting with objects in the environment, including their own bodies. The missing “third way”, namely working things out, can also involve exploratory activities, but the explorations can be done with *representations* of the objects and processes instead of using the actual physical objects. The representations used to explore possibilities can be entirely mental, e.g. visualising what happens when some geometrical configuration is transformed, or they can include diagrams or models, for instance 2-D pictures representing 3-D structures, with processes represented by marks on the pictures [23,33]. The biological advantages of being able to reason about future actions have often been pointed out, e.g. by [42,43].

Although reasoning with representations in place of the objects can be fallacious, and often is, nevertheless, when done rigorously, it is mathematical inference rather than empirical inference. As documented at length in [44] the methods of mathematics are far from infallible. But that does not make them empirical in the same way as the methods of the physical sciences are. This claim is subtle and complex and will not be substantiated here. A more detailed discussion can be found in the online presentation mentioned in Footnote 1.

The ability to think about and reason about novel combinations of familiar types of process is often required for solving new problems, for example realising for the first time that instead of going from A to B and then to C it is possible to take a short cut from A to C, or realising that a rigid circular disc can serve as well as something long and thin (like a screwdriver) to lever something up, or realising for the first time that a pair of long cylinders placed under a large box can make the box easier to move along a flat surface, though the cylinders will repeatedly have to be moved from the back of the pushed object to the front, as the motion proceeds.

The majority of work on reasoning in AI makes use of forms of representation that are close to logic and algebra, namely what were called "Fregean" representations in [23], because, as Frege was the first to note, they all use the function-argument syntactic form. However, humans, and possibly some other animals have the ability to visualise things changing and can use visualisation to work out consequences of processes. That was described as reasoning with "analogical" representations in the 1971 paper. [45] makes a very similar distinction, though he compares analogical representations with Fregean ones thus "What is a relation in one system may be a part, or an element, in the other."

5.9 Creative Reasoning About Processes and Affordances

[8] mention various kinds of "prospectivity" that develop in children but they focus only on empirically learnt kinds of predictive rules, and ignore the child's growing ability to design and represent novel complex multi-stage processes that can achieve some goal. Such prospectivity involves *working out* what will happen if something is done, instead of merely using a learnt correlation or imitating an observed action.

The ability to work things out is facilitated and enhanced by the ability to form verbal descriptions, as in inventing stories, but linguistic competence is not a prerequisite for the ability, as can be seen in the creative problem-solving of pre-verbal children and some other animals. Human children, and some other animals, seem to be able to work out the consequences of some actions using geometric and topological reasoning – an ability also required for doing some kinds of mathematics, e.g. proving theorems in Euclidean geometry. Before doing mathematics explicitly, children need to develop a kind of visual and manipulative fluency regarding spatial structures, relationships and processes that is built up by playing with many different examples. [33] present examples of activities that can help young children to develop their understanding of topological structures and processes. If we can produce a theory of what the information-processing mechanisms are that make that possible we may be in a much better position to design such educational games and toys.

As illustrated in Figure 6, affordances can interact in complex ways when combined, because of the changing spatial relationships of objects during the processes of performing the actions. A large chair may afford lifting and carrying

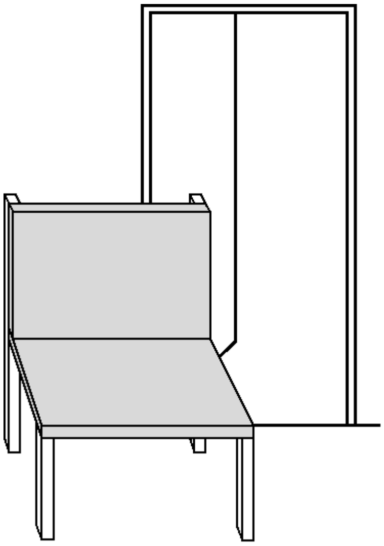

Fig. 6. *A person trying to move a chair that is too wide to fit through a door can work out how to move it through the door by combining a collection of translations and 3-D rotations about different axes, some done in parallel, some in sequence. Traditional AI planners cannot construct plans involving continuous interacting actions.*

from one place to another, and a doorway may afford passage from one room to another. But the attempt to combine the two affordances by lifting and carrying the chair to the next room may fail when the plan is tried, e.g. if it is found during the process of execution that the chair is too wide to fit through the doorway.

A very young child may not be able to do anything about that, but an older child who has learnt to perceive the possibility of rotation of a 3-D object, may realise that a combination of small rotations about different axes combined with small translations can form a compound process that results in the chair getting through the doorway. (Is any other type of animal capable of working that out?)

At an early stage the child may merely be able to do this one step at a time: seeing the possibility of the first rotation, then, after performing the rotation, seeing the possibility of forward motion, which is soon obstructed. Then another rotation may be tried followed by another translation to achieve the final goal. At a later stage the child may be able to see the possibility of the whole sequence of actions by visualising in advance the situation that will arise after each step. After searching in imagination through a set of possible action sequences the child may be able to work out by visual reasoning how to move the chair into the next room, and then do it. Although that description will be understood intuitively by most readers, it is not at all clear what sort of brain mechanism or computer mechanism can perform that reasoning function or achieve that learning.

If that is done often, the whole process may be learnt as a re-usable pattern for moving large objects that can be made to fit a variety of specific cases, without having to be re-discovered every time.

That learning requires some abstract, possibly parametrised, representation of the process to be created, stored in a re-usable form, and integrated with some kind of indexing mechanism that allows its relevance to be recognised when new related problems are encountered, so that the whole design process does not need to be repeated each time.

It is also possible to learn to separate complex affordances that have positive and negative aspects into components so as to retain only the positive aspects. This is a type of process "de-bugging" that children and adults often have to do, though many politicians ignore the need when designing policies. A colleague reported to me that his child had learnt that open drawers could be closed by pushing them shut. The easiest way to do that was to curl his fingers over the upper edge of the projecting drawer and push, as he would push other objects. The resulting pain led him to discover that the pushing could be made slightly less convenient by flattening his hand when pushing, so as to produce the desired result without the unwanted side effects.

It is important to notice the difference between merely discovering *empirically* that using a flat hand avoids the painful result (and achieves a completely shut drawer) and *seeing why* that is so. This requires understanding consequences of actions in which the gap between two surfaces is reduced while there is something between the surfaces.[10] Being able to do geometrical reasoning enables a child who is old enough to work out why pushing with a flat hand prevents fingers being caught between the two surfaces. Such a child could also work out that if the drawer is recessed and the front of the chest juts out immediately above the drawer, it will be necessary to make sure that the flat hand does not project above the edge of the drawer in the last stages of pushing.

5.10 The Need for Explanatory Theories

It is clear that many humans can perform such reasoning by visualising processes in advance of producing them, but it is not at all clear what representations are used to manipulate information about the shapes and affordances of the objects involved. A shallow answer (at least as old as Craik's book) is that we build internal models of the environment and manipulate them in order to discover the consequences. But this leaves open the question what such a model could be. If the model has too much in common with the things in the environment then the internal model cannot be part of the explanation of how the external processes are perceived and controlled.

[10] This is an example of Kantian causal reasoning discussed in presentations at WONAC, June 2007: http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#wonac

It is often thought that such a predictive model must be isomorphic with what it represents. An internal model that is *isomorphic* with the environment would lack appropriate explanatory power, since it would shift our problem to the problem of explaining how the person (or a "homunculus" perceiving the model), could work out what to do with it. Crude explanations in terms of internal models simply produce an infinite regress. The requirements for adequate theories are discussed in more detail in Section 6 below.

The lack of a suitable theory about how spatial structures and processes can be represented means that we cannot yet give a similar range of capabilities to robots, although a noteworthy early effort at giving a machine the ability to reason spatially (only in 2-D) about actions in the environment was Funt's PhD (1977).

Although there has been much work on giving machines with video cameras or laser scanners the ability to construct representations of 3-D structures and processes that can be projected onto a screen to show pictures or videos from different viewpoints, we still lack the ability to give robots spatial representation capabilities that are usable for other purposes such as manipulating objects, planning manipulations, and reasoning about them, except in very simple cases, for instance motions of objects that do not touch one another, or motions of objects that are all rigid and have simple shapes, e.g. moving cylinders or cubes.

AI planning systems developed in the 1960s as exemplified in the STRIPS planner [46] and more complex recent planners (surveyed in [47]) all make use of the fact that knowledge about affordances can be abstracted into reusable information about the preconditions and effects of actions. Once that is done, it provides a new kind of *cognitive* affordance: concerned with acting on information structures. That early AI work demonstrated the possibility of combining knowledge about simple actions to provide information about complex actions composed of simple actions.

However, STRIPS and its successors assume that the information about actions and affordances can be expressed in terms of implications between propositions expressed in a logical formalism. The planning process searches for a sequence (or partially ordered network) of discrete actions that will transform the initial problem state into the desired goal state. But we need a richer mechanism to handle actions that involve interactions between continuous processes, like the changes that occur while an arm-chair is being rotated and translated simultaneously, or while a sink is being wiped clean with a cloth.

5.11 Seeing Logical Relationships

Consider an argument like

```
All computers consume energy
Fred is using a computer,
Therefore
Fred is using something that consumes energy
```

When normally sighted people do logic or algebra they use the ability to *see* structural relationships between formulae: a kind of geometrical competence.

For example, detecting the logical validity of the above argument depends on noticing that there are structural relations between parts of the argument. People can learn to recognise such groups of sentences as instances of a spatial pattern, such as this:

All Ps are Qs
A is using a P
Therefore
A is using a Q

where "A", "P", and "Q" are variables capable of being replaced by a referring expression and two predicate expressions, respectively, to produce different instances. A still more general expression could be formed by replacing "is using" with a relation variable.

The examples in previous sections concerned acquiring and using information about what sorts of changes can and cannot occur in a physical situation. Something similar can occur when the ability is applied not to solid movable objects in the environment, but to patterns that can be perceived on the surfaces of objects, such as symbols drawn in the sand, or on paper representing logical structures.

Although humans normally check the structure of such reasoning using vision, there are other ways to check such an inference: a computer running the language Prolog may do it by performing the "resolution" operation on computer datastructures, i.e. matching two structures and linking variables in one to parts of the other. So the same reasoning process may be implemented in different sorts of mechanisms – including both visual and logical mechanisms.

At my first AI conference, in 1971, I challenged the then AI orthodoxy by arguing that intelligent machines would need to be able to reason geometrically as well as logically, and that some reasoning with diagrams should be regarded as being valid and rigorous, and in some cases far more efficient than reasoning using logic, because logical representations are topic-neutral and sometimes lose some of the domain structure that can be used in searching for proofs.

But it soon became clear that, although many people had independently concluded that AI techniques needed to be extended using spatial reasoning techniques, neither I nor anyone else knew how to design machines with the right kinds of abilities, even though there were many people working on giving machines the ability to recognise, analyse and manipulate images, or parts of images, often represented as 2-D rectangular arrays, though sometimes in other forms, e.g. using log-polar coordinates, e.g. [24]. Other examples were presented in [25]. More recent examples are [48,49].

5.12 An Objection: Blind Mathematicians

It could be argued that the description of mathematical reasoning as "visual" must be wrong because people who have been blind from birth can reason about shapes and do logic and mathematics even though they cannot see.[11] That argu-

[11] E.g. [50] reports on a number of blind mathematicians.

ment ignores the fact that some of the visual apparatus produced by evolution to support seeing and reasoning about structures and processes in the environment is in brain mechanisms that perform some of their functions without optical input: like the normal ability to see what *can* change in a situation when those changes are not occurring, or the ability to visualise a future sequence of actions that are not now being performed, and therefore cannot produce retinal input.

Moreover, since the same 3-D situation can generate infinitely many views, it would be explosively expensive to represent a recurring 3-D situation in terms of corresponding retinal contents, so it is possible that some biological organisms, including humans, have the ability to represent the environment using *amodal, non-retinotopic* forms of representation whose registration with the optic array changes across saccades and other physical movements (as discussed in [51,52]). Such a representation could also be linked to tactile and haptic information even if it originally evolved under pressure to cope with the vast amount of near and far scene information provided in parallel through vision. So people who have been blind from birth may still be using the bulk of the visual system that evolved in their ancestors, just as sighted people may be using it when they dream about seeing things, and when they visualise diagrams with their eyes shut. However, the process of learning about the specific spatial structures and processes in the environment must be very different for blind people. The learning processes are different in a different way for individuals born with other disabilities, e.g. with missing limbs, or with faulty physical control mechanisms as in cerebral palsy. The fact that many such individuals acquire a common humanity via different routes is an indication of how much of human mentality is independent of our specific form of embodiment. That is probably not true of all species, though many are highly adaptable if injured.

5.13 Use of Abstraction Is Not Metaphor

The claim that visual mechanisms using abstract patterns can support reasoning of the sort done in mathematics should not be confused with the common claim that spatial concepts are used as metaphors for non-spatial topics, for instance the claim that we must use spatial metaphors in thinking about numbers or about time. Such claims are based on a failure to understand that there are high level domain-neutral concepts (e.g. “order”, “between”, “more than”, “is a subset of”) which are *equally applicable* to many different domains for example because all those domains have some common topological features. Points on a line, times in a temporal interval, and any set of integers all form total orderings, and both can be divided into ordered subsets in many different ways. Seeing that there is an abstract, generally applicable, pattern, is different from seeing structural mappings between partly similar instances.

An animal or machine that can abstract the possibility of joining a vertex of a triangle to the middle of the opposite side (see Section 3.5) from instances of that process, and can form a pattern that is applicable to all triangles, past and future, is doing something different from creating a single such triangle with such a line drawn on it and mapping parts of it on to other triangles in order to modify them in the same way. Observing the commonality between modifications of two specific triangles may *trigger* the discovery of the general pattern, but if the general pattern has not been understood as a re-usable pattern with many instances, then each new instance of a triangle has to be separately tested for the possibility of being part of a process that can be mapped onto the original model.

Using manipulable structures in one domain to represent patterns in another domain for the purpose of reasoning about them, because both domains share some features (without necessarily being isomorphic), is a different matter from using the first domain as a metaphor for the second domain: metaphors do not provide valid inferences, although they are often usefully suggestive.

These ideas are not new: Several famous examples of visual proofs are presented in [53]. Many theorists, including great logicians such as Frege (see [54]) and mathematicians such as [55], have pointed at the use of visualisation and spatial reasoning capabilities in mathematics and logic. It will be clear from earlier comments about exosomatic ontologies and representations in Section 4 that I do not agree with Poincaré's claim "But every one knows that this perception of the third dimension reduces to a sense of the effort of accommodation which must be made, and to a sense of the convergence of the two eyes, that must take place in order to perceive an object, distinctly. These are muscular sensations quite different from the visual sensations which have given us the concept of the two first dimensions." I suspect he would have modified his views if he had been involved in designing robots that can perceive and reason about 3-D scenes.

However, I am not aware of work that spells out detailed engineering requirements for a working visual system capable of being used for mathematical visual reasoning, or work that proposes a design that can meet those requirements, although a few special cases have been partly modelled in AI programs, fairly recent examples being [48,49]. A partial set of requirements for such a system is in a presentation mentioned in Footnote 1.

6 Studying Mechanisms *vs.* Studying Requirements.

6.1 The Importance of Requirements

It has gradually become clear that finding suitable explanatory mechanisms is only one part of the problem of studying minds, including studying vision. Less obvious than the need to find mechanisms is the need to clarify precisely what the mechanisms are needed for. It is not very difficult to specify hundreds of different algorithms for analysing, comparing or transforming images or parts of images represented as 2-D arrays – as was done, for example, in the 1960s, in Azriel Rosenfeld's research group in Maryland [56][12] – but not so easy to specify what the high level requirements are that determine what sorts of algorithm

[12] It is worth remembering that in those days computers ran millions of times more slowly than now, and memories were measured in kilobytes not gigabytes, so that things that are trivial now were monumental achievements then.

and what forms of representation are needed for modelling or replicating human visual abilities, including the ability to reason visually.

Engineers are accustomed to distinguishing specifying *requirements* from specifying *a design*. The requirements can guide the search for designs, in addition to providing criteria for evaluating designs. However, we still have no comprehensive generally agreed inventory of the capabilities that need to be modelled and explained in an artificial, human-like visual system, although there are many fragmentary requirements studied by different researchers in psychology, neuroscience, AI, education, history of art [57], etc. This paper adds more fragments to the collection but does not claim completeness.

Unfortunately, AI researchers (and other modellers) too often launch into seeking designs, on the assumption that the requirements are clear, e.g. because they think everyone knows what a visual system has to be able to do, or because they deliberately focus on a very narrow set of requirements defined by behaviours observed in psychology experiments, or requirements defined by some benchmark test, such as recognition of objects in a collection of images. There is nothing intrinsically wrong with such research, but it can lead researchers (and their students) to ignore the question of what else needs to be explained in addition to those behaviours. Moreover, as already remarked, successful models or explanations of a limited set of behaviours may not scale out.

A deeper problem, is that there is as yet not even a generally agreed ontology for discussing requirements and designs. That is we do not have an agreed set of concepts for describing cognitive functions in great detail, with sufficient precision to be used in specifying requirements for testable working systems. (For example, how would you decide whether a robot really is visualising a route, or merely constructing and manipulating a data-structure representing that route?)

6.2 Mistaken Requirements

Section 5 of [34], entitled “Previous false starts”, presents a list of nine fairly common mistaken assumptions about how vision systems should work. Much of this paper is implicitly an extension of that list. For example it is tempting to suppose that a requirement for 3-D vision mechanisms is that they must construct a 3-D model of the perceived scene, with the components arranged within the model isomorphically with the relationships in the scene. This is in fact how some AI vision systems work, for instance in robots that move around using laser range-finders to create a model of all the visible surfaces in the environment. Such a model can be used to generate graphical displays showing the appearance of the environment from different viewpoints. However, Figure 7 shows that we are able to see a complex structure without building such a model, for there cannot be a model of an impossible scene.

An alternative hypothesis is that what a visual system needs to do is construct a large collection of fragments of information of various sorts about surfaces, objects, relationships, possible changes and constraints on changes in the scene (including positive and negative affordances), with most of the information fragments represented at least approximately in registration with the optic

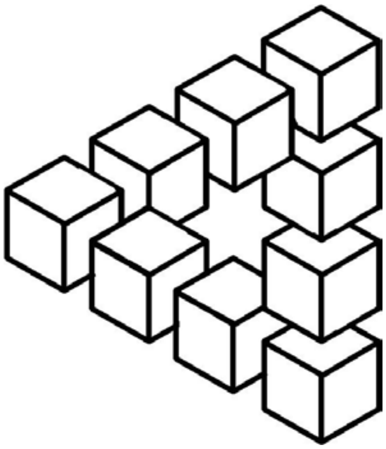

Fig. 7. *This figure (after a drawing by Oscar Reutersvärd in 1934) has many components that can be interpreted as representing parts of a 3-D scene, with a wide variety of affordances, e.g. possible ways of moving the individual cubes, or ways of inserting something into the gaps between the cubes. Yet the whole scene, made up of all the fragments with the depicted relationships and locally consistent affordances, is geometrically impossible. This shows that seeing the scene depicted here cannot involve constructing a model isomorphic with the whole scene, since no such model can exist.*

array, though in an amodal form, insofar as they refer to entities that are not all currently visible, or use a 3-D ontology and have the potential to be linked to control of actions or to be matched against different sorts of sensory information. This form of representation, which could include spatial GLs (Sec. 2.2), can be thought of as a generalisation of generalised aspect-graphs.

The idea of an aspect-graph has probably been reinvented several times using different labels. For instance, in [58] they were called "frame systems". The core idea is that a 3-D object will present distinct 2-D views which can be linked to form a graph where the edges of the graph represent actions that a viewer can perform, such as moving left, or right or up or down. These actions are associated with both continuous changes in the 2-D view (e.g. relative lengths of lines and sizes of angles changing) and also discontinuous changes, e.g. an edge or face disappearing or coming into view. This seems to be the same idea as Kant discussed in [2] in connection with different views presented by the same house as you move around it or move up or down within it. This idea can be generalised so that more actions are included, such as touching or pushing, or grasping an object and more changes are produced such as two objects coming together or moving apart, or an object rotating, or sliding or tilting, or becoming unstable, etc,

An animal or robot that does not have a good representation of the 3-D structures, relationships and processes involved will have to build up many generalised aspect graphs empirically, using myriad viewpoints, different ways of performing the same action, etc., in order to learn all the relevant mappings between sensorimotor signals and the resulting sensorimotor signals. However if a representation of the relevant 3-D structures and processes is available, along with mechanisms that are able to work out geometric and topological consequences of changing relationships, that can be used to derive consequences that have never been experienced before, e.g. the result of performing an action with your left hand on a green triangular block of a particular size for the first time. For this reason, we need to understand the differences between *somatic* and *exosomatic* ontologies and representations, where the former are concerned with patterns (at various levels of abstraction) in sets of sensory and motor signals and the latter are concerned with entities in the environment that exist independently of anything perceiving or acting on them. Somatic, sensorimotor ontologies refer to things that can only exist in the body of an animal or robot. A great deal of current research in vision and robotics focuses on mechanisms that manipulate only representations of sensorimotor phenomena, e.g. statistical patterns relating multi-modal sensor and motor signals, whereas one of the main claims of this paper is that human-like systems need, in addition, amodal exosomatic ontologies and forms of representation suited to them. Their great advantage is that a single representation of a process can in the environment, such as two fingers grasping a berry, can ignore all the variations in sensor and motor signals that depend on precisely how the grasping is done and the viewpoint from which the process is observed, and which other objects may partially occlude relevant surfaces.

6.3 Is Consistency-checking Required?

Of course, this ability to acquire such general, reusable forms of representation, requires the perceiver to be able to take in visual (and possibly haptic and tactile) sensory input and construct a representation of the 3-D structure in the scene (as opposed to only representations of the appearances). If that representation of the 3-D scene is made up of many piecemeal representations of fragments of the scene and the possible effects of processes involving those fragments, then in principle those fragments could form an inconsistent totality as shown in Figure 7. So it would seem that an intelligent robot or animal must constantly check whether it has consistent percepts.

However, since no portion of the 3-D environment is capable of containing impossible objects, there is normally no need for such a visual system to check that all the derived information is consistent, except in order to eliminate ambiguities. This is just as well since in general consistency checking is an intractable process, which scales exponentially with the number of items to be checked.

Humans can learn to check consistency, at least in some contexts, and that enables them to see that the configuration of cubes depicted in the figure is impossible. However a very young child will not notice the impossibility, and even an adult might not notice the impossibility if a picture of an impossible scene contains a large number of objects arranged not in an triangle but in a more complex configuration.

6.4 Obvious and Unobvious Requirements: Ontological Blindness

Many people, e.g. [59], have noticed the need for hierarchical decomposition of complex perceived objects and the usefulness of a mixture of top down, bottom up and middle out processing in perception of such objects. A recent example of a visual learning system that automatically acquires a layered network of image features as a result of being exposed to a collection of pictures showing objects that the system learns to recognise is [60].

What is not so often noticed is that in addition to *part-whole* hierarchies there are also *ontological* layers, as illustrated in the Popeye program[13] described in Chapter 9 of [32]. The program was presented with images made of dots in a rectangular grid, such as Figure 8, which it analysed and interpreted in terms of:

- a layer of dot configurations (which could, for example, contain collections of collinear adjacent dots);
- a layer of line configurations, where lines are interpretations of "noisy" sets of collinear dots, and can form configurations such parallel pairs, and junctions of various sorts;
- a layer of 2-D overlapping, opaque 'plates' with straight sides and rectangular corners, which in Popeye were restricted to the shapes of cut-out capital letters, such as "A", "E", "F", "H" etc. represented in a noisy fashion by collections of straight line segments;
- a layer of sequences of capital letters represented by the plates, also in a "noisy" fashion because the plates could be jumbled together with overlaps;
- a layer of words, represented by the letter sequences.

The Popeye program illustrated the need for some visual systems to use ontologies at different levels of abstraction processed concurrently, using a mixture of top-down, bottom-up and middle-out processing, where lower levels are not *parts* of the higher levels but rather *represent* the higher levels.

At each ontological layer there are part-whole hierarchies. E.g. a complex group of dots may be made of smaller groups of dots. Going from one ontological layer to another is not a matter of grouping parts into a whole, but *interpreting* one sort of structure as representing another, for instance, interpreting configurations of dots as representing configurations lines. A complex configuration of lines could be made of simpler configurations, which are ultimately made of line-segments. Configurations of lines can be interpreted as representing overlapping 2-D plates. A complex opaque plate (e.g. the plate representing the "E" in Figure 8) could be made of smaller, simpler rectangular plates. In this situation

[13] So-called because it was implemented in the Edinburgh University AI programming language POP-2.

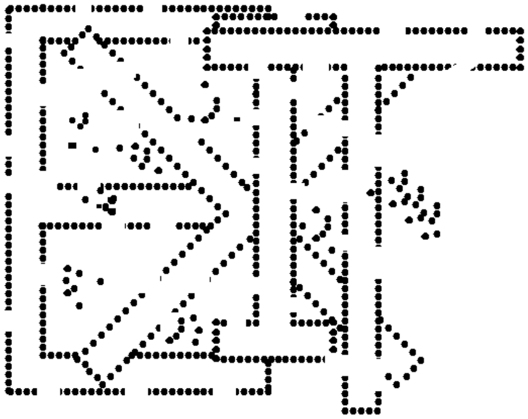

Fig. 8. *This is a typical example of a configuration of dots presented to the Popeye program in Chapter 9 of Sloman (1978), which attempted to find a known word by concurrently looking for structures in several ontological layers, with a mixture of top-down, bottom-up and middle-out influences. If the noise and clutter were not too bad, the program, like humans could detect the word before identifying all the letters and their parts. It also degraded gracefully as noise and clutter made the problem harder.*

a dot may be a part of a row of dots, but it is not a part of a plate or letter or a word, though it is part of something that may *represent* plates letters or words.

The same letters and words can be represented in different ways, e.g. using different fonts, and different conventions for projecting fonts into 2-D configurations. For example, in one place a word may be represented using outline letters, and in another place using filled letters, though Popeye could not handle that. Popeye could handle overlapping letters, though most other text-reading programs do not allow the notion of overlap and would fail on these pictures, unlike humans.

Text represented using several ontological layers may be regarded as a very contrived example, but similar comments about ontological layers can be made when a working machine is perceived, such as the internals of an old fashioned clock. There will be sensory layers concerned with changing patterns of varying complexity in the optic array. A perceiver will have to interpret those changing sensory patterns as representing 3-D surfaces and their relationships, some of which change over time. At a higher level of abstraction there are functional categories of objects, e.g. levers, gears, pulleys, axles, strings, and various more or less complex clusters of such objects, such as escapement mechanisms.

The ability to perceive the operation of a complex machine may have to use an ontology including different kinds of substance, for instance both rigid parts, flexible elastic strings or springs, flexible chains with weights on the end, whose density needs to be high. At a still higher level of abstraction there are causal and functional roles, such as providing energy to drive the machine, transmitting energy from one part to another, and in some kinds of engine mechanisms for varying the torque and controlling speed either by modifying the energy source or by applying brakes, etc.

There are many vision researchers who appreciate the need for a vision system to move between a 2-D ontology and a 3-D ontology. For a recent survey see [61]. The need for such layers will be evident to anyone who works on vision-based text-understanding. However it is rare to include as many ontological categories, in different layers as I claim are needed by an intelligent human-like agent interacting with a 3-D environment, or to relate those layers to different processing layers in the central architecture as explained in [62].

6.5 Different Uses of 3-D Information

Another subtle issue is the need to contrast merely requiring information about 3-D structure (e.g. "amodal completion" of perceived volumes) with specifying that the information needs to be represented so that it can be used in a certain way. Representation of information for the purposes of recognition involves different requirements from the representations that can be used for projecting images from different viewpoints, for servo-control of manipulation, for planning future actions, or for understanding how something works. The need for representations to be usable for diverse applications is another example of the problem of "scaling out", mentioned in Section 3.

Perception of intelligent agents in the environment involves yet another level of abstraction, insofar as some perceived movements are interpreted as actions with purposes. For instance a hand moving towards a cup might be seen as intentional, whereas changing patterns of wrinkles on a sleeve, or motion of shadows might be seen as unintended side-effects.

6.6 Seeing Mental States

Moreover, if eyes and face are visible, humans will often see not just actions but also mental states, such as focus of attention in a certain direction, puzzlement, worry, relief, happiness, sadness, and so on. Insofar as these are all *seen* rather than inferred in some non-visual formalism, the percepts will be at least approximately in registration with the optic array. Happiness is seen in one face and not in another. The requirement for perceptual mechanisms to use an ontological layer that includes mental states raises many problems that will not be discussed here, for example the need to be able to cope with referential opacity. Representing something that is itself an information user requires meta-semantic competences. These subtleties are ignored by researchers who train computer programs to label pictures of faces using words such as "happy", "angry", and claim that their programs can recognise emotional states.

The ability to perceive some processes as intentional actions produced by other agents with mental states, including desires and beliefs probably had to evolve (and has to develop in children or robots) before the ability to produce and understand intentional communications.

The need for ontological layers to be used in perceptual processing was noticed long ago in connection with natural language understanding, which involves, for example, phonemic, morphemic, syntactic, semantic and pragmatic layers. So, since one of the uses of vision is reading written language, it should have been obvious that visual perception also requires ontological layers: but that fact has been generally ignored by vision researchers, except in those cases where the distinction between the 2-D ontology of images and the 3-D ontology of scenes has been noticed. The point being made here is that those two ontological layers do not suffice for the full variety of human and animal visual perception. We need to add several more layers including: 2-D and 3-D processes, causal relations, functional relations, actions, mental states of perceived agents, social phenomena, and all the forms of perception that differ from one adult specialisation to another, e.g. in athletics, hunting animals, various kinds of craft, engineering, scientific research, etc.

Figure 8 shows that even within the set of 2-D phenomena that play a role in visual perception, there are different ontological levels, that are relevant to different cognitive sub-functions. For instance, painters, unlike sculptors, have to learn to ignore what they know about the 3-D structures they see and attend to the 2-D relations features and relationships to be depicted in drawings and paintings, some related to edges, some to surface markings, some to shading, some to surface curvature, and so on.

6.7 Perceiving 2-D Processes in the Optic Array

2-D processes involving changes in the optic array are also important, as J.J. Gibson pointed out. As noted in [16], apart from perception of static scenes, vision is also required for online control of continuous actions (visual servoing) which requires different forms of representation from those required for perception of structures to be described, remembered, used in planning actions, etc.

Sometimes a 2-D projection is more useful than a 3-D description for the control problem, as it may be simpler and quicker to compute, and can suffice for particular task, such as steering a vehicle through a gap.

But it is a mistake to think that only continuously varying percepts are involved in online visual control of actions: there is also checking whether goals or sub-goals have been achieved, whether the conditions for future processes have not been violated, whether new obstacles or new opportunities have turned up, and so on. Those can involve checking discrete conditions.

Unfortunately research on ventral and dorsal streams of neural processing has led some researchers (e.g. [63]) to assume that control of action is separate from cognition, or worse, that spatial perception ("where things are") is a completely separate function from categorisation ("what things are"), apparently ignoring the fact that what an object is may depend on where its parts are in relation to one another, or where it is located in a larger whole.

6.8 Structures and Processes

A major theme that pervades engineering design and especially software design is the relationship between structure and process. But the more general relevance of this theme is not always noticed. For instance, doing school Euclidean geometry involves seeing how a particular structure can be the starting point for various processes, and seeing how processes of construction can produce new structures from old ones, e.g. in proving theorems, such as Pythagoras' theorem.

Some processes transform structures discretely, e.g. by changing the topology of something (adding a new line to a diagram, separating two parts of an object, altering contact or containment relations) others continuously (e.g. painting a wall, pushing or lifting an object, or blowing up a balloon).

Understanding how an old-fashioned clock works involves seeing causal connections and constraints related to possible processes that can occur in the mechanism.

Performing many actions involves doing several things concurrently, e.g. (a) producing processes (e.g. grasping), (b) seeing those processes, and (c) using visual servoing to control the fine details, (d) predicting future processes. Some or all of this may be done unconsciously, as in posture control and many skilled performances. (Such unconscious use of expertise does not make the actions unintentional.)

Another theme that has been evident for many decades is the fact that percepts can involve hierarchical structure, although not all the structures should be thought of as loop-free trees like parse-trees. Seeing a bicycle, or even a simple closed polygon, requires use of a graph rather than a tree, though to a first approximation most animals and plants have a tree-like structure (e.g. decomposition into parts that are decomposed into parts, etc.)

6.9 Layered Ontologies

We have seen, in Section 6.4, that in addition to part-whole decomposition, perception can use layered ontologies. For example, one sub-ontology might consist entirely of 2-D image structures and processes, whereas another includes 3-D spatial structures and processes, and another kinds of 'stuff' of which objects are made and their properties (e.g. rigidity, elasticity, solubility, thermal conductivity, etc.), to which can be added mental states and processes, e.g. seeing a person as happy or sad, or as intently watching a crawling insect.

The use of multiple ontologies is even more obvious when what is seen is text, or sheet music, perceived using different geometric, syntactic, and semantic ontologies.

What did not strike me until 2005 when I was working on an EU-funded robot project (CoSy) is what follows from the combination of the two themes (a) the content of what is seen is often processes and process-related affordances, and (b) the content of what is seen involves both hierarchical structure and multiple ontologies. These themes together imply a set of requirements for a visual system that makes current working models seem very far from what we need either in order to understand human and animal vision, or in order to produce working models for scientific or engineering purposes.

Very many people are now working on how to cope with the fact that digital camera technology, occlusion of one object by another, poor lighting, confusing colours and textures, intervening fog or dirty windows, and other common occurrences leads to pervasive problems of noise and ambiguity that have to be accounted for. This has led to a lot of research on mechanisms for representing and manipulating uncertainty, for instance propagating inferences based on noisy and ambiguous low level information.

What is not always noticed is that a consequence of use of layered ontologies is that humans have ways of seeing high level structures and processes whose descriptions are impervious to the low level uncertainties: you can see that there definitely is a person walking away from you on your side of the road and another walking in the opposite direction on the other side of the road, even though you cannot tell the precise locations, velocities, accelerations, sizes, orientations and other features of the people and their hands, feet, arms, legs, etc. The latter information may be totally irrelevant for your current purposes (looking to see whether any cars are coming, before you cross the road).

Is it possible that the processes of dealing with uncertainty could be made far more efficient and much simpler if they were ignored for a while and some effort put into the problem of finding the determinate, certain, higher level information first and then adding the uncertain details constrained by what is already certain? Of course, finding ontologies, forms of representation. and mechanisms to perform those high level tasks may be very difficult.

Some notes on this can be found in this discussion paper on predicting affordance changes, including both action affordances and epistemic affordances: http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0702

6.10 Seeing is Prior to Recognising

Much research on visual perception considers only one of the functions of perception, namely recognition. It is often forgotten that there are many things we can see that we cannot recognise or label, and indeed that is a precondition for learning to categorise things.

When you see a complex new object that you do not recognise you may see a great deal of 3-D structure, which includes recognising many types of *surface fragment*, including flat parts, curved parts, changes of curvature, bumps, ridges, grooves, holes, discontinuous curves where two curved parts meet, and many more. In addition to many surface fragments, many of their relationships are also seen.

Not only are relationships within objects important, but also relationships between objects, and also between different parts of objects. When someone is seen to grasp something manually, there are many different configurations that can be involved in grasping the same object, depending on how different parts of the hand are related to different parts of the grasped object. For many years my own work on vision (e.g in the Popeye project) assumed that perception involves perception of structure at different levels of abstraction, although the need to perceive and control *processes* during continuous visual servoing was mentioned in [16]. However, it was not until working on requirements for a robot that can manipulate 3-D objects in 2005 that I realised that it is also necessary to be able to see *processes* at different levels of abstraction, as explained with a number of examples in this online presentation: `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0505`.

6.11 Seeing Processes

Biological considerations suggest that, for most animals, perception of processes must be the most important function, since perception is crucial to the control of action, in a dynamic, sometimes rapidly changing environment that can include mobile predators and mobile prey, and where different parts of the environment provide different nutrients, shelter, etc. So from this viewpoint perception of structures is just a special case of perception of processes – processes in which not much happens.

Unfortunately, not only has very little (as far as I know) been achieved in designing visual systems that can perceive a wide range of 3-D spatial structures (as opposed to recognising objects in images), there is even less AI work on perception of processes, apart from things like online control of simple movements which involves sensing one or two changing values and sending out simple control signals, for instance “pole balancing” control systems. There seems also to be very little research in psychology and neuroscience on the forms of representations and mechanisms required for perception of processes involving moving or changing structures, apart from research that merely finds out who can do what under what conditions. Examples of the latter include [64], [65] and [66].

Addressing that deficiency (including explaining how GLs for process representation work) should be a major goal for future vision research, both in computational modelling but also in neuroscience. Some speculations about mechanisms are presented in Section 7.

6.12 Seeing Possible Processes: Proto-affordances

We have already noted that an important feature of process perception is the ability to consider different ways a process may continue, some of them conditional on other processes intervening, such as an obstacle being moved onto or off the path of a moving object. Many cases of predictive control include some element of uncertainty based on imprecise measurements of position, velocity or acceleration. This sort of uncertainty can be handled using fuzzy or probabilistic control devices which handle intervals instead of point values.

However there are cases where the issue is not uncertainty or inaccuracy of measurement but the existence of very different opportunities, such as getting past an obstacle by climbing over it, or going round it on the left or on the right. It may be very clear what the alternatives are, and what their advantages and disadvantages are. E.g one alternative may involve a climb that requires finding something to stand on, while another requires a heavy object to be pushed out of the way, and the third requires squeezing through a narrow gap.

The ability to notice and evaluate distinct possible futures is required not only when an animal is controlling its own actions but also when it perceives something else whose motion could continue in different ways. How the ability to detect such vicarious affordances is used may depend on whether the perceived mover is someone (or something) the perceiver is trying to help, or a prey animal, or a predator.

Groups concerned with attempting to conserve an endangered species sometimes have to learn to recognise affordances for members of that species.

In simple cases, prediction and evaluation of alternative futures can make use of a simulation mechanism. But the requirement to deal explicitly with alternative possibilities requires a more sophisticated simulation than is needed for prediction: a predictive simulation can simply be run to derive a result, whereas evaluation of alternatives requires the ability to start the simulation with different initial conditions so that it produces different results. It also requires some way of recording the different results so that they can be used later for evaluation or further processing. For example, it may be necessary to use the fact that after the first choice new situations can arise with new choices that depend on the first choice. In relatively simple domains, such as discrete board games, storing multiple branching futures going several steps ahead may use fairly simple logical or other forms of representation (though space requirements can expand exponentially with number of steps considered, so that early pruning of poor alternatives is usually required).

The ability to cope with branching futures in a continuous spatial environment poses problems that do not arise in “toy” discrete grid-based environments. The agent has to be able to chunk continuous ranges of options into relatively small sets of alternatives in order to avoid dealing with explosively branching paths into the future. How to do this may be something learnt by exploring good ways to group options by representing situations and possible motions at a high level of abstraction. For example all the motions that share some topological feature such as entering a certain region of space can often be grouped together as one option.

Learning to see good ways of subdividing continuous spatial regions and continuous ranges of future actions involves developing a good descriptive ontology at a higher level of abstraction than sensor and motor signals inherently provide. The structure of the environment, not some feature of sensorimotor signals makes it sensible to distinguish the three cases: moving forward to one side of an obstacle, moving so as to make contact with the obstacle and moving so as to go to the other side. Further useful subdivisions may also be generated by the environment, e.g. if the wall beyond the left side of the obstacle has two doors known to lead into the same room beyond the wall, and only the further door is open.

In addition to “chunking” of possibilities on the basis of differences between opportunities for an animal or robot to move as a whole there are ways of chunking them on the basis of articulation of the agent's body into independently movable parts. For example, if there are two hands available, and some task requires both hands to be used, one to pick an object up and the other to perform some action on it (e.g. removing its lid) then each hand can be considered for each task, producing four possible combinations. However if it is difficult or impossible for either hand to do both tasks, then detecting that difficulty in advance may make it clear that the set of futures should be pruned by requiring each hand to do only one task, leaving only two options. Noticing that one task is better suited to one of the hands can then reduce the set under consideration to one case, even though that case covers a very large variety of slightly different processes in the space of motion trajectories.

In humans, and some other species, during the first few years of life a major function of play and exploration in an infant is providing opportunities for the discovery of many hundreds of concepts that are useful for chunking sets of possible states of affairs and possible process, and learning good ways to represent them so as to facilitate predicting high level consequences, which can then be used in rapid decision-making strategies.

Being able to detect, reason about, compare and evaluate alternative possible futures requires a form of representation that goes beyond mere simulation of the motion, although a simulation that can be restarted at saved decision points and pushed in different directions when restarted could play a role – and that may have been a precursor to more sophisticated forms of planning.

More sophisticated mechanisms are required if the results of different forward projections need to be stored and compared.[14] That requires use of a form of representation that can express abstract summaries of the alternatives, whose relationships can also be represented during a comparison process, before a decision is taken, whether it is a decision about what to predict or a decision about what to do. In the 1960s and 1970s AI researchers showed how to do some of this using logic-based forms of representation. Doing it with spatial GLs (Sect. 2.2) could use partly similar mechanisms, but different forms of representation would be required.

The ability to perceive not just what is happening at any time but what the possible branching futures are – including, good futures, neutral futures, and bad futures from the point of view of the perceiver's goals and actions, is an aspect of J.J. Gibson's theory of perception as being primarily about *affordances for the perceiver* rather than acquisition of information about some objective and neutral environment. However, I don't think Gibson considered the need to be able to represent, compare and evaluate multi-step branching futures: that would have been incompatible with his adamant denial of any role for representations and computation.

[14] E.g. using a "fully deliberative" architecture, defined in `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604`

7 Speculation About Mechanisms Required: New Kinds of Dynamical System

Preceding sections have assembled many facts about animal and human vision that help to constrain both theories of how brains, or the virtual machines implemented on brains work, and computer-based models that are intended to replicate or explain human competences. One thing that has not been mentioned so far is the extraordinary speed with which animal vision operates. This is a requirement for fast moving animals whose environment can change rapidly (including animals that fly through tree-tops). An informal demonstration of the speed with which we can process a series of unrelated photographs and extract quite abstract information about them is available online here http://www.cs.bham.ac.uk/research/projects/cogaff/misc/multipic-challenge.pdf Compare people turning corners, looking out of trains, coming out of railway stations or airports in new towns, watching TV documentaries about places never visited, etc.

No known mechanism comes anywhere near explaining how that is possible especially at the speed with which we do it.

7.1 Sketch of a Possible Mechanism

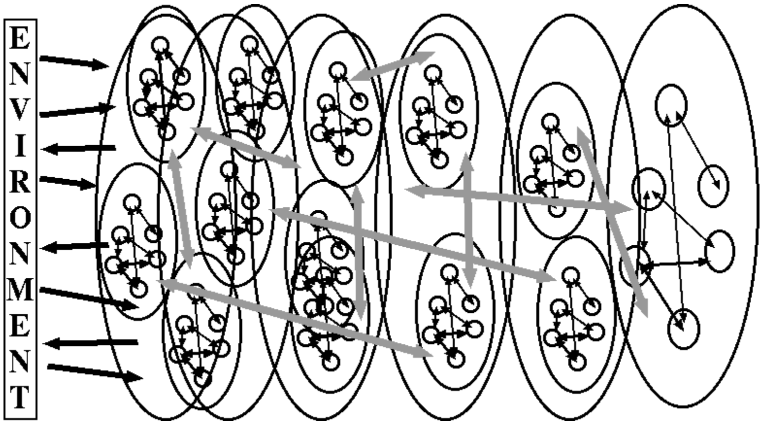

Fig. 9. *A crude impressionistic sketch indicating a collection of dynamical systems some closely coupled with the environment through sensors and effectors others more remote, with many causal linkages between different subsystems, many of which will be dormant at any time. Some of the larger dynamical systems are composed of smaller ones. The system does not all exist at birth but is grown, through a lengthy process of learning and development partly driven by the environment, as sketched in Chappell and Sloman 2007*

Perhaps we need a new kind of dynamical system. Some current researchers (e.g., [68]) investigate cognition based on dynamical systems composed of simple "brains" closely coupled with the environment through sensors and effectors. We need to extend those ideas to allow a multitude of interacting dynamical systems, some of which can run decoupled from the environment, for instance during planning and reasoning, as indicated crudely in Figure 9. During process perception, changing sensory information will drive a collection of linked processes at different levels of abstraction. Some of the same processes may occur when possible but non-existent processes are imagined in order to reason about their consequences.

Many dynamical systems are defined in terms of continuously changing variables and interactions defined by differential equations, whereas our previous discussion, e.g. in Section 5.6, implies that we need mechanisms that can represent discontinuous as well as continuous changes, for example to cope with topological changes that occur as objects are moved, or goals become satisfied. Another piece of evidence for such a requirement is the sort of discrete 'flip' that can occur when viewing well known ambiguous figures such as the Necker cube, the duck-rabbit, and the old-woman/young-woman picture. It is significant that such internal flips can occur without any change in sensory input.

It is possible that adult human perception depends on the prior construction of a very large number of multi-stable dynamical systems each made of many components that are themselves made of "lower level" multistable dynamical systems. Many of the subsystems will be dormant at any time, but the mechanisms must support rapidly activating an organised, layered, collection of them partly under the control of current sensory input, partly under control of current goals, needs, or expectations, and partly under the control of a large collection of constraints and preferences linking the different dynamical systems.

On this model, each new perceived scene triggers the activation of a collection of dynamical systems driven by the low level contents of the optic array and these in turn trigger the activation of successively higher level dynamical systems corresponding to more and more complex ontologies, where the construction process is constrained simultaneously by general knowledge, the current data, and, in some cases, immediate contextual knowledge. Sub-systems that are activated can also influence and help to constrain the activating subsystems, influencing grouping, thresholding, and removing ambiguities, as happened in the Popeye program described in Section 6.4.

As processes occur in the scene or the perceiver moves, that will drive changes in some of the lower level subsystems which in turn will cause changes elsewhere, causing the perceived processes to be represented by internal processes at different levels of abstraction. Some the same mechanisms may be used when when possible but non-existent processes are imagined in order to reason about their consequences.

On this view, a human-like visual system is a very complex multi-stable dynamical system:

– composed of multiple smaller multi-stable dynamical systems

- that are grown over many years of learning,
- that may be (recursively?) composed of smaller collections of multi-stable dynamical systems that can be turned on and off as needed,
- some with only discrete attractors, others capable of changing continuously,
- many of them inert or disabled most of the time, but capable of being activated rapidly,
- each capable of being influenced by other sub-systems or sensory input or changing current goals, i.e. turned on, then kicked into new (more stable) states bottom up, top down or sideways,
- constrained in parallel by many other multi-stable sub-systems,
- with mechanisms for interpreting configurations of subsystem-states as representing scene structures and affordances, and interpreting changing configurations as representing processes,
- using different such representations at different levels of abstraction changing on different time scales,
- where the whole system is capable of growing new sub-systems, permanent or temporary, some short-term (for the current environment) and some long term (when learning to perceive new things), e.g.
  - learning to read text
  - learning to sight read music
  - learning to play tennis expertly,
    etc.

That specification contrasts with “atomic-state dynamical systems”, described in [69] as dynamical systems:

- with a fixed number of variables that change continuously
- with one global state
- that can only be in one attractor at a time
- with a fixed structure (e.g. a fixed size state vector).

The difficulties of implementing a dynamical system with the desired properties (including components in which spatial GLs are manipulated) should not be underestimated. The mechanisms used by brains for this purpose may turn out to be very different from mechanisms already discovered.

8 Concluding Comments

In [34] it was proposed that we need to replace ‘modular’ architectures with ‘labyrinthine’ architectures, reflecting both the variety of components required within a visual system and the varieties of interconnectivity between visual subsystems and other subsystems (e.g. action control subsystems, auditory subsystems, and various kinds of central systems).

One way to make progress may be to start by relating human vision to the many evolutionary precursors, including vision in other animals. If newer systems did not replace older ones, but built on them, that suggests that many research questions need to be rephrased to assume that many different kinds of visual processing are going on concurrently, especially when a process is perceived that involves different levels of abstraction perceived concurrently, e.g. continuous physical and geometric changes relating parts of visible surfaces and spaces at the lowest level, discrete changes, including topological and causal changes at a higher level, and in some cases intentional actions, successes, failures, near misses, etc. at a still more abstract level. The different levels use different ontologies, different forms of representation, and probably different mechanisms, yet they are all interconnected, and all in partial registration with the optic array (not with retinal images, since perceived processes survive saccades).

It is very important to take account of the fact that those ontologies are not to be defined only in terms of what is going on inside the organism (i.e. in the nervous system and the body) since a great deal of the information an organism needs is not about what is happening in it, but what is happening in the environment, though the environment is not some unique given (as implicitly assumed in Marr's theory of vision (1982), for example) but is different for different organisms, even when located in the same place. They have different niches.

As Ulric Neisser pointed out in his (1976) it is folly to study only minds and brains without studying the environments those minds and brains evolved to function in.

One of the major points emphasised here is that coping with our environment requires humans to be able to perceive, predict, plan, explain, reason about, and control processes of many kinds, and some of that ability is closely related to our ability to do mathematical reasoning about geometric and topological structures and processes. So perhaps trying to model the development of a mathematician able to do spatial reasoning will turn out to provide a major stepping stone to explaining how human vision works and producing convincing working models. Perhaps it will show that Immanuel Kant got something right about the nature of mathematical knowledge, all those years ago.

9 Acknowledgements

Many of the ideas reported in this paper were developed as part of the requirements analysis activities in the EU funded CoSy project www.cognitivesystems.org. I am especially indebted to Jeremy Wyatt and other members of the Birmingham CoSy team. Mohan Sridharan kindly read and commented on a draft. Jackie Chappell has helped with biological evidence, the work on nature-nurture tradeoffs and the difference between Humean and Kantian causal understanding. Dima Damen helped me improve Figure 9. Shimon Edelman made several very useful comments.

References

1. Sloman, A.: Knowing and Understanding: Relations between meaning and truth, meaning and necessary truth, meaning and syn-

thetic necessary truth. PhD thesis, Oxford University (1962) http://www.cs.bham.ac.uk/research/projects/cogaff/07.html#706.

2. Kant, I.: Critique of Pure Reason. Macmillan, London (1781) Translated (1929) by Norman Kemp Smith.
3. Sloman, A.: Kantian Philosophy of Mathematics and Young Robots. To appear in proceedings MKM08 COSY-TR-0802, School of Computer Science, University of Birmingham, UK (March 2008) http://www.cs.bham.ac.uk/research/projects/cosy/papers#tr0802.
4. Kaneff, S., ed.: Picture language machines. Academic Press, New York (1970)
5. Marr, D.: Vision. Freeman, San Francisco (1982)
6. Berthoz, A.: The Brain's sense of movement. Perspectives in Cognitive Science. Harvard University Press, London, UK (2000)
7. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston, MA (1979)
8. Gibson, E.J., Pick, A.D.: An Ecological Approach to Perceptual Learning and Development. Oxford University Press, New York (2000)
9. McDermott, D.: Artificial intelligence meets natural stupidity. In Haugeland, J., ed.: Mind Design. MIT Press, Cambridge, MA (1981)
10. Sloman, A.: Actual possibilities. In Aiello, L., Shapiro, S., eds.: Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96), Boston, MA, Morgan Kaufmann Publishers (1996) 627–638
11. Ryle, G.: The Concept of Mind. Hutchinson, London (1949)
12. Clark, A.: Contemporary Problems in the Philosophy of Perception: A review prompted by *The Contents of Experience: Essays on Perception*, Ed. Tim Crane. Cambridge: Cambridge University Press, 1992. American Journal of Psychology **107**(4 Winter 1994) (1994) 613–22 available at http://selfpace.uconn.edu/paper/aconline.htm.
13. Sloman, A.: Two Notions Contrasted: 'Logical Geography' and 'Logical Topography' (Variations on a theme by Gilbert Ryle: The logical topography of 'Logical Geography'.). Technical Report COSY-DP-0703, School of Computer Science, University of Birmingham,, Birmingham, UK (2007) http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0703.
14. Sloman, A.: Architecture-based conceptions of mind. In: In the Scope of Logic, Methodology, and Philosophy of Science (Vol II). Synthese Library Vol. 316. Kluwer, Dordrecht (2002) 403–427 http://www.cs.bham.ac.uk/research/projects/cogaff/00-02.html#57.
15. Sloman, A.: Putting the Pieces Together Again. In Sun, R., ed.: Cambridge Handbook on Computational Psychology. Cambridge University Press, New York (2008) http://www.cs.bham.ac.uk/research/projects/cogaff/07.html#710.
16. Sloman, A.: Image interpretation: The way ahead? In Braddick, O., Sleigh., A., eds.: Physical and Biological Processing of Images (Proceedings of an international symposium organised by The Rank Prize Funds, London, 1982.). Springer-Verlag, Berlin (1982) 380–401 http://www.cs.bham.ac.uk/research/projects/cogaff/06.html#0604.
17. Sloman, A.: The structure of the space of possible minds. In Torrance, S., ed.: The Mind and the Machine: philosophical aspects of Artificial Intelligence. Ellis Horwood, Chichester (1984)
18. Sloman, A.: Explorations in design space. In Cohn, A., ed.: Proceedings 11th European Conference on AI, Amsterdam, August 1994, Chichester, John Wiley (1994) 578–582

19. Sloman, A.: Exploring design space and niche space. In: Proceedings 5th Scandinavian Conference on AI, Trondheim, Amsterdam, IOS Press (1995)
20. Sloman, A.: Interacting trajectories in design space and niche space: A philosopher speculates about evolution. In M.Schoenauer, *et al.*., ed.: Parallel Problem Solving from Nature – PPSN VI. Lecture Notes in Computer Science, No 1917, Berlin, Springer-Verlag (2000) 3–16
21. Sloman, A.: The primacy of non-communicative language. In MacCafferty, M., Gray, K., eds.: The analysis of Meaning: Informatics 5 Proceedings ASLIB/BCS Conference, Oxford, March 1979, London, Aslib (1979) 1–15 http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#43.
22. Sloman, A., Chappell, J.: Computational Cognitive Epigenetics (Commentary on [67]). Behavioral and Brain Sciences **30**(4) (2007) 375–6 http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0703.
23. Sloman, A.: Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In: Proc 2nd IJCAI, London, William Kaufmann (1971) 209–226 http://www.cs.bham.ac.uk/research/cogaff/04.html#200407.
24. Funt, B.V.: Whisper: A problem-solving system utilizing diagrams and a parallel processing retina. In: IJCAI, Cambridge, MA, IJCAI'77 (1977) 459–464
25. Glasgow, J., Narayanan, H., Chandrasekaran, B., eds.: Diagrammatic Reasoning: Computational and Cognitive Perspectives. MIT Press, Cambridge, MA (1995)
26. Sloman, A.: Diversity of Developmental Trajectories in Natural and Artificial Intelligence. In Morrison, C.T., Oates, T.T., eds.: Computational Approaches to Representation Change during Learning and Development. AAAI Fall Symposium 2007, Technical Report FS-07-03, Menlo Park, CA, AAAI Press (2007) 70–79 http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0704.
27. Sloman, A., Chappell, J.: The Altricial-Precocial Spectrum for Robots. In: Proceedings IJCAI'05, Edinburgh, IJCAI (2005) 1187–1192 http://www.cs.bham.ac.uk/research/cogaff/05.html#200502.
28. Chappell, J., Sloman, A.: Natural and artificial meta-configured altricial information-processing systems. International Journal of Unconventional Computing **3**(3) (2007) 211–239 http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609.
29. Lakatos, I.: The methodology of scientific research programmes. In Worrall, J., Currie, G., eds.: Philosophical papers, Vol I. Cambridge University Press, Cambridge (1980)
30. Sloman, A.: 'Necessary', 'A Priori' and 'Analytic'. Analysis **26**(1) (1965) 12–16 Now online http://www.cs.bham.ac.uk/research/projects/cogaff/07.html#701.
31. Whitehead, A.N., Russell, B.: Principia Mathematica Vols I – III. Cambridge University Press, Cambridge (1910–1913)
32. Sloman, A.: The Computer Revolution in Philosophy. Harvester Press (and Humanities Press), Hassocks, Sussex (1978) http://www.cs.bham.ac.uk/research/cogaff/crp.
33. Sauvy, J., Suavy, S.: The Child's Discovery of Space: From hopscotch to mazes – an introduction to intuitive topology. Penguin Education, Harmondsworth (1974) Translated from the French by Pam Wells.
34. Sloman, A.: On designing a visual system (towards a gibsonian computational model of vision). Journal of Experimental and Theoretical AI **1**(4) (1989) 289–337 http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#7.
35. Ellis, R., Tucker, M.: Micro-affordance : The potentiation of components of action by seen objects. British journal of psychology **91**(4) (2000) 451–471

36. Warneken, F., Tomasello, M.: Altruistic helping in human infants and young chimpanzees. Science (3 March 2006) 1301–1303 DOI:10.1126/science.1121448.
37. Weir, A.A.S., Chappell, J., Kacelnik, A.: Shaping of hooks in New Caledonian crows. Science **297** (2002) 981
38. Hayworth, K.J., Biederman, I.: Neural evidence for intermediate representations in object recognition. Vision Research (In Press) **46**(23) (2006) 4024–4031 doi:10.1016/j.visres.2006.07.015.
39. Marr, D.: Analysis of occluding contour. Proc. Roy. Soc. Lond. **B197** (1977) 441–475
40. Varley, P.A.C., Martin, R.R., Suzuki, H.: Can Machines Interpret Line Drawings? In Hughes, J.F., Jorge, J.A., eds.: Sketch-Based Interfaces and Modelling Eurographics Symposium Proceedings. Eurographics Association (2004) 107–116 http://ralph.cs.cf.ac.uk/papers/Geometry/Can.pdf.
41. Miller, G., Galanter, E., Pribram, K.: Plans and the Structure of Behaviour. Holt, New York (1960)
42. Craik, K.: The Nature of Explanation. Cambridge University Press, London, New York (1943)
43. Popper, K.: Objective Knowledge. Oxford University Press, Oxford (1972)
44. Lakatos, I.: Proofs and Refutations. Cambridge University Press, Cambridge, UK (1976)
45. Attneave, F.: How Do You Know? The American Psychologist **29** (1974) 493–499
46. Fikes, R., Nilsson, N.: STRIPS: A new approach to the application of theorem proving to problem solving. Artificial Intelligence **2** (1971) 189–208.
47. Ghallab, M., Nau, D., Traverso, P.: Automated Planning, Theory and Practice. Elsevier, Morgan Kaufmann Publishers, San Francisco, CA (2004)
48. Jamnik, M., Bundy, A., Green, I.: On automating diagrammatic proofs of arithmetic arguments. Journal of Logic, Language and Information **8**(3) (1999) 297–321
49. Winterstein, D.: Using Diagrammatic Reasoning for Theorem Proving in a Continuous Domain. PhD thesis, University of Edinburgh, School of Informatics. (2005) http://www.era.lib.ed.ac.uk/handle/1842/642.
50. Jackson, A.: The World of Blind Mathematicians. Notices of the American Mathematical Society **49**(10) (2002) http://www.ams.org/notices/200210/comm-morin.pdf.
51. Trehub, A.: The Cognitive Brain. MIT Press, Cambridge, MA (1991) http://www.people.umass.edu/trehub/.
52. Grush, R.: The emulation theory of representation: Motor control, imagery, and perception. Behavioral and Brain Sciences **27** (2004) 377–442
53. Nelsen, R.B.: Proofs without words: Exercises in Visual Thinking. Mathematical Association of America, Washington DC (1993)
54. Merrick, T.: What Frege Meant When He Said: Kant is Right about Geometry. Philosophia Mathematica **14**(1) (2006) 44–75 doi:10.1093/philmat/nkj013.
55. Poincaré, H.: Science and hypothesis. W. Scott, London (1905) http://www.archive.org/details/scienceandhypoth00poinuoft.
56. Rosenfeld, A.: Picture Processing by Computer. Academic Press, New York (1969)
57. Gombrich, E.H.: Art and Illusion: A Study in the Psychology of Pictorial Representation. Pantheon, New York (1960)
58. Minsky, M.L.: A framework for representing knowledge. In Winston, P.H., ed.: The psychology of computer vision. McGraw-Hill, New York (1978) 211–277
59. Neisser, U.: Cognitive Psychology. Appleton-Century-Crofts, New York (1967)

60. Fidler, S., Leonardis, A.: Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts. In: Proceedings Conference on Computer Vision and Pattern Recognition,, Minneapolis, IEEE Computer Society (2007) 1–8 http://vicos.fri.uni-lj.si/data/alesl/cvpr07fidler.pdf.
61. Breckon, T.P., Fisher, R.B.: Amodal volume completion: 3D visual completion. Computer Vision and Image Understanding (99) (2005) 499–526 doi:10.1016/j.cviu.2005.05.002.
62. Sloman, A.: Evolvable biologically plausible visual architectures. In Cootes, T., Taylor, C., eds.: Proceedings of British Machine Vision Conference, Manchester, BMVA (2001) 313–322
63. Goodale, M., Milner, A.: Separate visual pathways for perception and action. Trends in Neurosciences **15**(1) (1992) 20–25
64. Heider, F., Simmel, M.: An experimental study of apparent behaviour. American Journal of Psychology **57** (1944.) 243–259
65. Michotte, A.: The perception of causality. Methuen, Andover, MA (1962)
66. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception and Psychophysics **14** (1973) 201–211
67. Jablonka, E., Lamb, M.J.: Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. MIT Press, Cambridge MA (2005)
68. Beer, R.: Dynamical approaches to cognitive science. Trends in Cognitive Sciences **4**(3) (2000) 91–99 (http://vorlon.case.edu/ beer/Papers/TICS.pdf).
69. Sloman, A.: The mind as a control system. In Hookway, C., Peterson, D., eds.: Philosophy and the Cognitive Sciences. Cambridge University Press, Cambridge, UK (1993) 69–110 http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#18.
70. Neisser, U.: Cognition and Reality. W. H. Freeman., San Francisco (1976)