# Varieties of Metacognition in Natural and Artificial Systems

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

### Abstract

Some AI researchers aim to make useful machines, including robots. Others aim to understand general principles of information-processing machines whether natural or artificial, often with special emphasis on humans and human-like systems: They primarily address scientific and philosophical questions rather than practical goals. However, the tasks required to pursue scientific and engineering goals overlap considerably, since both involve building working systems to test ideas and demonstrate results, and the conceptual frameworks and development tools needed for both overlap. This paper, partly based on requirements analysis in the CoSy robotics project, surveys varieties of meta-cognition and draws attention to some types that appear to play a role in intelligent biological individuals (e.g. humans) and which could also help with practical engineering goals, but seem not to have been noticed by most researchers in the field. There are important implications for architectures and representations.

## Varieties of Requirements and Designs

The workshop manifesto (Cox and Raja 2007) states *"The 21st century is experiencing a renewed interest in an old idea within artificial intelligence that goes to the heart of what it means to be both human and intelligent. This idea is that much can be gained by thinking of one's own thinking. Metareasoning is the process of reasoning about reasoning itself."* This implies that the idea is not restricted to engineering goals, but includes the scientific and philosophical study of humans. As is clear from Cox (2005) the scientific concern with metacognition in AI goes back to the founders. Scientific and philosophical aims have always been my primary reason for interest in AI, including meta-cognitive mechanisms (e.g. in chapters 6 and 10 of my 1978). Study of other animals should also be included, since humans are products of a process that produced organisms with many sizes, shapes, habitats, competences, and social organisations; and we cannot expect to understand all the design tradeoffs in humans unless we compare alternatives.

Such comparisons could also be of great importance for biology/ethology. That would involve studying both the space of requirements (*niche* space) and the space of designs that can be assessed against those requirements (*design* space). Assessment need not be production of a *measurement*, e.g. a number or a total 'fitness' ordering. Instead comparisons could produce structured descriptions of strengths and weaknesses in various conditions and in relation to various functions (like the more useful consumer reports and Minsky (1963)).

One way to do that comparative study is to attempt analytically to retrace the steps of biological evolution. Simply simulating evolution does not necessarily yield any new understanding of design-options or tradeoffs, however impressive the end-products. But retrospective analysis does not need to follow the chronology of evolution: working backward analytically may be as informative as working forward, in studying both evolution and individual development. Since the whole evolutionary process was so long, so broad (because many things evolved in parallel) and so intricate, it may be most fruitful to attempt to identify major design discontinuities, producing before-after comparisons of both requirements and designs and analysing their implications, both for science (including psychology, biology, neuroscience) and for engineering. A partial, purely illustrative, survey of this sort was presented in Sloman (2007a).

Philosophy, especially conceptual analysis, will inevitably be involved in the process. This paper attempts to identify issues to be addressed in such analytical comparative investigations, starting from a collection of design features of humans that are not widely recognized.

## Beyond the Manifesto

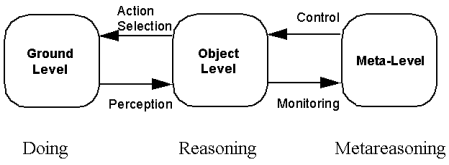


Figure 1: *From the workshop manifesto*

One implication of the generalisation to biological phenomena (and human-like robots) is that the "ground level" referred to in the manifesto (Fig. 1) may include arbitrarily complex physical and social environments. In humans, while awake, there are sensors and effectors continuously

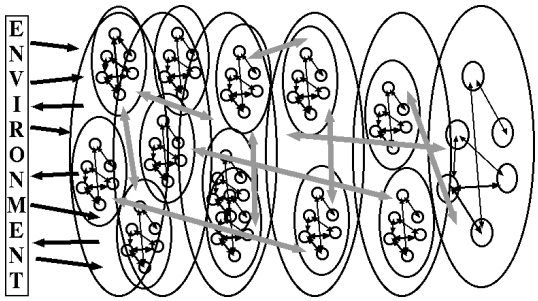


Figure 2: *Dynamical subsystems vary in many ways including degree of environmental coupling, speed of change, whether continuous or discrete, what is represented, etc.*

coupled to the environment: i.e. the coupling does not alternate between being on and off while more central processes analyse sensory inputs or decide what to do. Consequently, instead of an "action-perception cycle", we need an architecture with *concurrent* processes of many kinds, which can interact with one another. (Even a single-cpu computer can support concurrent enduring processes because, while the cpu is shared, the majority of the state of each process endures in memory. However, some kinds of concurrency may require specialised hardware, e.g. for continuous control.)

So the arrows, instead of representing cyclic transitions between states represented by the boxes, as in flow charts, must represent flow of information and control between *enduring* sub-systems operating at different levels of abstraction, on different time-scales, some changing continuously, others discretely, and performing different functions (as in Figs 2 and 3). This has deep implications for forms of representation, algorithms, and architectures, and for possible interactions and conflicts between sub-systems.

Such concurrency was out of the question in the early days of AI: computers were far too slow and had far too little memory. If finding the rim of a teacup in an image takes 20 minutes, a robot cannot perceive and act concurrently. There are also application domains where continuous monitoring and control are out of the question because everything is discrete and all objects are static, e.g. most of the internet.

## Control Hierarchies

The manifesto states that *"Much of the research is driven by the problems of limited rationality"*, but there is a much older, more general requirement, namely the well-known requirement for hierarchical control. It seems clear that that requirement was "discovered" millions of years ago by evolution and addressed in a wide variety of organisms. Instead of designing a control mechanism so that it deals with all circumstances that can arise, which will make it very complex and non-modular, it is often better (and much simpler) to add another mechanism that monitors the performance of the control mechanism and on the basis of what is detected modifies the behaviour, either on the fly by changing parameters perhaps, or by altering the module concerned so as to modify future behaviours – as happens in many learning systems and self-debugging systems. An early AI example of

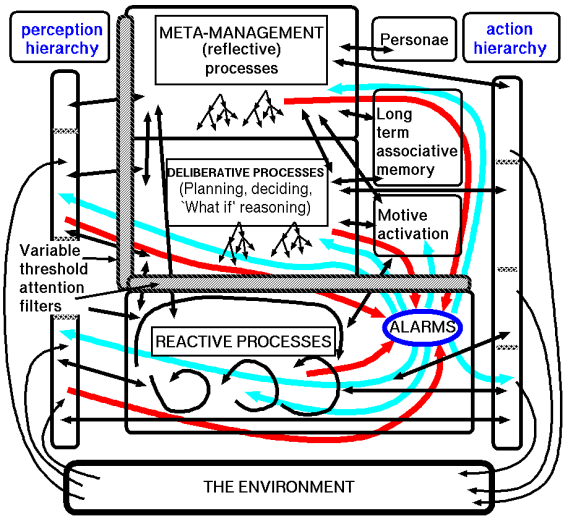


Figure 3: *A sketchy representation of a human-like architecture specification, H-CogAff, developed within the CogAff project. Alarm mechanisms can produce different types of episodic emotions. The architecture grows itself.*

this was Sussman's HACKER (1975). A more recent example, using concurrent control at different levels of abstraction is Brooks' subsumption architecture (1986). Compare the A, B, and C brains in Minsky (1987), and Minsky (2006).

Another obvious benefit of layered control for a system (or subsystem) that needs to be able to function in many different contexts, is that instead of a monolithic design to cope with all contexts, it may be much simpler to have different modules each suited to a specific class of situations, and a meta-level module that monitors changing contexts and decides (possibly after learning) when to switch control between modules. Otherwise each module would frequently have to invoke code for testing whether it should continue, and if not which other module to transfer control to, a more complex, wasteful, and error-prone design.

A dedicated meta-level module may be able to improve specific modules separately without risk of unwanted interactions. It can also detect that no existing module is adequate to a certain task, and produce a new one with appropriate applicability conditions, possibly by copying and editing one or more of the older ones, or by using planning mechanisms to create a new module for the new context.

Meta-level decisions may themselves involve arbitrarily complex problems and their switching systems may also be monitored and modulated by higher-level controllers. In principle, such a control philosophy can involve arbitrarily many layers of meta-control, but in practice there will be limits (Minsky 1968).

## "Alarm" mechanisms

Another design option is to include a type of "alarm" module (Fig. 3) that is always on, and normally does nothing but monitor processes (possibly both external and internal) but which is capable of rapidly detecting situations that need

emergency control actions, possibly involving modifying the behaviour of large numbers of other modules, for instance producing states such as freezing (in order to avoid detection or to attend to a potential threat), fleeing, feeding, fighting or mating (the "five Fs" – freezing is often omitted). Other options in emergencies include: slowing down, changing direction, invoking special perceptual capabilities, doing more exhaustive analysis of options, etc. Some alarm mechanisms performing these functions need to act very quickly to replace or modulate current active modules, so they will need fast pattern recognition mechanisms rather than reasoning processes. Several "emotional" states can be defined in terms of different patterns of activity resulting from monitoring and modulating done by such alarm mechanisms in a layered control hierarchy (Sloman 2001).

Environments change, and neither evolution nor human designers can anticipate all possible developments, so it is important for such alarm mechanisms to be trainable. By analysing different sorts of training, e.g. using immediate feedback, long term feedback, or more sophisticated debugging, we can distinguish different kinds of architecture, and different forms of "emotional learning" that can occur in those architectures.

## Meta-management

Following Beaudoin (1994), we emphasise the variety of types of "meta-" level functioning by using the label "meta-management" (Fig 3), suggesting control as well as monitoring, reasoning, learning, etc. Switching modules and debugging are examples already mentioned.

If alarm mechanisms use fast, possibly crude, pattern matching, they can produce false positives and false negatives, though training may reduce both. So another meta-management control function is to determine when alarm systems should be allowed to interrupt ongoing processes, possibly by varying attention filter thresholds, or by other means, illustrated in the work of Beaudoin and Wright (1977). As shown in Wright *et al* (1996), some potentially disruptive control states can endure in a dormant, but easily reawakened mode, over long periods, e.g. in grief, infatuation, obsessive ambition, jealousy, etc., contradicting the common view of emotions as necessarily episodic, expressed behaviourally and entering into consciousness. For more on the architectural basis of diverse affective states and processes see Sloman, Chrisley, and Scheutz (2005).

Learning not to react in some situations that produce fright or avoidance is one kind of emotional learning. Another is learning how to modulate the "disruptive" control mechanisms so as to minimise harmful consequences, e.g. when controlling a dangerous vehicle or machine.

Architectures with different meta-management functions can support different types of mental state (Sloman 2002). In humans, these higher level layers seem to develop over many years, changing the mental architecture, including developing high level control-regimes that could be labelled different "Personae" (Fig 3 top right). Some researchers unfortunately *define* "architecture" as something fixed, which rules out consideration of this sort of development over time, e.g. Langley and Laird (2006).

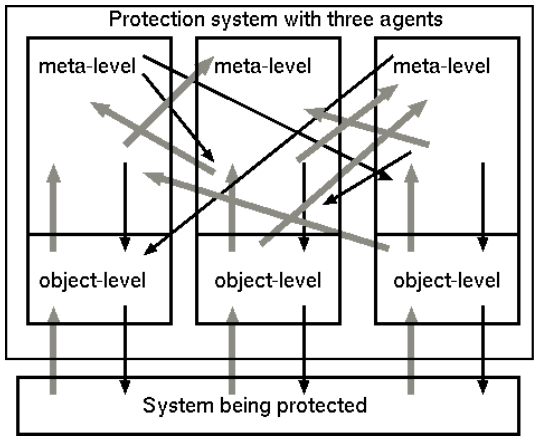


Figure 4: *Catriona Kennedy's PhD showed how a software protection system could use "mutual meta-management" to improve resistance to attack. Upward pointing grey arrows represent monitoring, and downward pointing black arrows represent acting. Not all possible arrows are shown.*

## Non-hierarchical Control

Fig 4 illustrates the possibility of several meta-management subsystems all monitoring the same lower level system (e.g. watching an operating system for signs of intrusion) and simultaneously monitoring one another. This idea, partly inspired by animal immune systems, was developed by Kennedy to demonstrate ways to make intrusion detection systems more robust (Kennedy and Sloman 2003). After training, each monitor can detect anomalies in its own behaviour, the behaviour of any of the other monitors, or in the behaviour of the main system, and can repair the anomaly by restoring the original code to the defective module. So an intruder must simultaneously disable all the monitors to avoid detection. There is no claim that this kind of system occurs in nature: it is merely mentioned here as a possible design that might be useful in some artificial systems. However if telepathy were possible, it might allow a group of humans to enter into a mutual meta-management pact!

## Tools

This work, and other work done within our group since about 1994, made use of the SimAgent toolkit[1] which was designed to allow exploration of various kinds of architectures with concurrently active components of different sorts, e.g. reactive, deliberative and meta-management capabilities, with different working memories for different subsystems. Unlike most others (e.g. ACT-R, SOAR, etc.) it was not designed to support a particular type of architecture.

An unusual feature is use of rules as communication channels: conditions of a rule can use one working memory and actions of the same rule another. Moreover, since a rule can have conditions checking several different memories and actions altering different memories, a ruleset can have some

[1] http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html. This could be used for work in robotics, but in order to support C++ and Java a new toolkit has been developed Hawes et al. (2007) and Hawes, Zillich, and Wyatt (2007)

of the features of a neural net, blurring the neural/symbolic boundary. This is crucial for the implementation of meta-management (monitoring and control). And since the rules are themselves editable entities in changeable memories, a working system can, in principle, grow its own architecture. However, experiments using that possibility have not yet been attempted apart from Kennedy's work.

This flexibility also allows exploration of ways in which architecture construction can go wrong. This might provide a basis for exploring various neural dysfunctions, including autism, and some aspects of schizophrenia. For example, mechanisms for determining which internal states and processes have normal internal causes and which are externally caused might go wrong, leading to internally caused events (e.g. thoughts occurring) being classified as externally caused, as happens in schizophrenia. A theoretical analysis of possible meta-management dysfunctions could be a contribution to clinical psychology suggesting new empirical observations.

## Meta-management and Consciousness

It is often suggested that consciousness depends on the existence of something like a meta-management layer in an architecture, though details differ (Minsky 1968; Sloman 1978; Johnson-Laird 1988; Baars 1988; Shanahan 2006). However the concept of "consciousness" (like "emotion") is so riddled with confusion and muddle that it is best not treated as part of a well defined requirement for AI, or being used to specify any clear scientific research question.

But there are clearer special cases. For instance, the notion of an individual having self-awareness is easier to discuss (McCarthy 1995). It covers a wide variety of cases, including things other individuals could know about the individual ("external self-consciousness"), e.g. knowing where you are, knowing what you have just done, knowing that you are about to do something risky, knowing that you are smiling, knowing how you look to others, knowing that you are offending someone, and so on. Self-consciousness also includes various types of introspection ("internal self-consciousness"), some trivial, such as a program checking the contents of a register, and some non-trivial e.g. an architecture that includes one or more self-observation subsystems running concurrently with others and using a "meta-semantic" ontology (defined below) that refers to relatively high level (e.g. representational) states, events and processes in the system, expressed in enduring multi-purpose forms of representation, as opposed to transient, low-level contents of conditionals and selection procedures, (Sloman 2007b).

A system using non-trivial introspection to acquire information about its internal states and processes, including possibly intermediate data-structures in perceptual and motor sub-systems, could be said to be self-aware, or to have self-consciousness of a sort. I think this subsumes cases discussed in McCarthy (1995), and also much of what philosophers say about "qualia" and "phenomenal consciousness".

Introspection is a kind of perception and any perception mechanism has the potential for error, notwithstanding the philosophically seductive Cartesian claim that knowledge of how things seem to you is infallible (Schwitzgebel 2007). That claim, "I cannot be mistaken about how things *seem* to me", or "I cannot be mistaken about the contents of my own experience", needs to be recognised as a trivial but confusing tautology, like "a voltmeter cannot be mistaken about what voltage it reports". What seems to you to be going on inside you cannot be different from what seems to you to be going on inside you, but it may be different from what is actually going on inside you. Intelligent reflective robots may fall into the same confusion.

## Meta-semantic Competence

Every control system that acquires and uses information has semantic competence, whether (like most neural nets) it is restricted to information expressed in changing scalar parameters or (like symbolic AI systems) it can handle structural information about states of affairs and processes involving more or less complex objects with parts, features, and static and changing relationships. There are many forms of representation for such information, including logics, natural language sentences, differential equations, databases, maps, and pictures, though computers do not (yet) use as many as humans do.

*Meta-semantic competence* is needed when a system uses information about information, or information about things that acquire, derive, use, contain or express information.

That requires the ability to represent things that represent, and to represent what they represent. This extension to semantic competence involves representing things like beliefs, goals, and plans that do not have the normal functions of beliefs, goals and plans, because they are not the beliefs, goals or plans of the system itself. More subtly, an individual $A$ with meta-semantic competence may need to represent information $I$ in another individual $B$ where the content $I$ has presuppositions that $A$ knows are false, but $B$ does not.

For example, $B$ may think there are fairies in his garden and have the goal of keeping them happy. $A$ must be able to represent the content of $B$'s beliefs and goals even though $A$ does not believe there are such fairies. Further, $A$ may know that a description D1 refers to the same thing as description D2, and therefore can substitute D2 for D2 in various representing contexts. But if $B$ does not know the equivalence, such substitutions may lead to mistaken conclusions about $B$'s beliefs and goals. These are the problems of "referential opacity", which contrasts with the "referential transparency" of simpler forms of representation.

Various philosophers and logicians have made proposals for dealing with these problems by adding new logical operators to standard logic, producing modal belief logics for example. An alternative approach, instead of using *notational* extensions, is to design intelligent systems with *architectural* extensions that allow information to be represented in special "encapsulated" modes, that prevent normal uses of the information. This is useful for story telling, for searching for explanations of observed events, for planning possible future actions, for counter-factual reasoning, for formulating questions to be answered, and also for representing information about how things were in the past, or will be in some future state, or how they are in some other location.

Such uses of encapsulation can arise even if there are no other information-users in the environment.

If an architecture provides support for encapsulation, that mechanism or a similar mechanism can be used for various meta-semantic purposes, such as representing mental states or information contents of other things. An example of such a mechanism is the ATT-META system of Barnden (http://www.cs.bham.ac.uk/~jab/ATT-Meta/) which handles encapsulation along with uncertainty. It also supports metaphorical reasoning.

Another use of meta-semantic competence is, of course, representation of information about one's own information processing. When representing what one previously thought, intended, or wanted, or what one *would have* thought, intended or wanted in a different situation, the issues of referential opacity are not very different from those involved in dealing with information states of others. Whether those problems carry over to first person present cases is debatable: e.g. some would argue that there is no significant difference in communicative content between the firm assertion "X is P" and the firm assertion "I believe that X is P". I shall not discuss that issue now.

Other topics for investigation include which animals have which sorts of meta-semantic competence, when and how such competences develop in young children, what changes in the child's brain and mind to make that possible, and how such abilities evolved. A more detailed investigation would show what sorts of meta-semantic competence are required for the meta-management architectural layer in Fig 3, and for the higher level visual capabilities required for seeing someone as happy, sad, puzzled, looking for something, etc.

## Ontology Development

Intelligent systems may be born or designed with the ontologies they need to use in perceiving and categorising things (a feature of *precocial* biological species), or, as in some altricial species Sloman and Chappell (2005), may have to develop their own ontologies through exploration and experiment, using mechanisms that evolved to support self-extension driven by interaction with an environment containing 3-D structures and processes as well as other intelligent individuals. Chappell and Sloman (2007) refer to largely genetically determined competences as "preconfigured" and those produced by layers of acquired competences as "meta-configured".

In the latter case the genome or initial design may be specifically suited to learning and development in a 3-D spatial and social environment but not to a *specific* 3-D social environment. Layered development processes can start by learning from the environment how to learn more in that environment. That can include learning what one can do and what sorts of consequences follow: a kind of meta-learning. This relates to epistemic affordances discussed below. How the genome can provide for such layered learning, including development of layers of ontology is not yet understood.

One form of ontology development is part of the process of construction of an explanatory theory. For example, an individual exploring the environment may find inexplicable differences between the behaviours of some of the things in the environment and postulate that the differences are caused by unobservable properties of the things, e.g. different kinds of material. This is analogous to theory formation and conceptual extension in science, and is part of the explanation of why symbol-grounding theory is mistaken, as explained in http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#models.

## Meta-management and theory-development

One of the uses of a meta-management capability is discovering the need to modify or extend current theories about how things work in the environment. as a result of noticing that some things perceived in the environment are puzzling because they conflict with predictions that are made on the basis of existing theories.

Sometimes such anomalies can be explained by a process of abduction, leading to a new theory making use of old concepts to state how behaviours of objects depend on observable conditions, e.g. explaining why a beam does not balance in the middle of its length if its thickness varies along its length.

If old concepts do not suffice, the new theory has to use new concepts referring to the unobserved but hypothesised properties that explain behaviour, for example the hypothesised property we call "magnetism". Unfortunately, the search space for abduction of new theories is explosively expanded if additional undefined symbols can be introduced.

So it may be important for the learner to use meta-management capabilities to identify features of what is and is not known, that can guide the creation of new concepts. I am not aware of any work in AI that has managed to model this process of simultaneous extension of ontology and theory, but if scientists and young children can do it it must be mechanisable. (Some aspects of the process were discussed in Chapter 2 of Sloman (1978).)

Ontology development is needed not only for perception of things in the environment, but also for internal self-monitoring and other meta-management uses – i.e. extending the individual's meta-semantic competences.

Here too, the new concepts may either be definable in terms of old ones, or may substantially extend the ontology, as happened in Freud's theories. An interesting case would be an individual A coming to the conclusion that another individual B had a sense that A lacked, because A can confirm that B has an unexplained way of gaining information that A can check is correct. For colour-blind people this is a common experience.

Some kinds of meta-semantic ontology extension may result from self-organising capabilities of self-monitoring mechanisms, for example an introspective mechanism that develops an ontology for describing intermediate states in perceptual processing, e.g. describing tastes, colour sensations, kinds of process, or functions of objects. When that happens some of the categories developed for use in internal monitoring (i.e. recording information about sensory contents) may be in principle uncommunicable to others because the concepts are 'causally indexical', i.e. implicitly referring to the mechanisms that do the classification, as suggested in Sloman and Chrisley (2003). This could be the source of some philosophical puzzles about qualia.

## A Meta-Turing Test for Theories of Mentality

At present the scope for theories of meta-cognition, including meta-management, seems to be very unconstrained. A possible constraint is suggested in Sloman (2007b), namely: an adequate theory of human meta-cognition should be capable of explaining how different thinkers with the same architecture can reach opposed views on many philosophical problems about mind, e.g. about the nature of human consciousness, human free will (Sloman 1992), human emotional states, and other controversial philosophical topics.

## Affordances and proto-affordances

There is a type of meta-cognition associated with perception of the environment, that requires Gibson's 1979 notion of affordance to be generalised. Many researchers, e.g. Marr (1982), assume that the function of vision (or more generally perception) is to provide information about geometrical and physical properties and relations of objects in the environment, including, for example, orientation, curvature, texture and colour of their visible surfaces, their distances, their spatial relationships, and their motion. Against this, Gibson argued that such information is not necessarily of use to organisms: they need, instead, information about which of the actions they are capable of are doable in a particular situation, and which ones will or will not produce desired results. I.e. he thought that perception provided information about positive and negative action affordances for the perceiver.

Without endorsing Gibson's bizarre explanations of how the information was acquired (using "direct pickup" or "resonance"), Sloman (2008a) treats the Gibsonian revolution as being of profound importance, though we need to go a long way beyond Gibson along the same road.

In particular, the perception of affordances that are related to possible actions depends on a more fundamental type of perception of "proto-affordances", namely possible processes and constraints on processes involving motion of 3-D objects and object fragments (e.g. possible changes in relations between portions of surfaces). Information about such proto-affordances can be acquired independently of whether the processes are produced by, or can be produced by, the perceiver or any other agent, and independently of whether they are relevant to the perceiver's needs or goals. An example might be seeing how a branch can move in the breeze and how other branches constrain its motion.

## Manipulation, Meta-cognition Mathematics

Not only can we perceive proto-affordances, we can also *reason* about their interactions when processes occur in close spatial proximity, e.g. working out consequences of making a new machine with interconnected pulleys, ropes, chains, levers, springs and gear wheels (Sloman 1971).

For example, if one end of a long, straight, rigid object is moved down while the centre is fixed, the other end *must* move up. At first the learner might discover such facts as mere statistical correlation. Later, reflection on what is understood by the assumption of rigidity, namely that there is some unknown feature of the internal structure of the material that prevents change of shape, can lead to the realisation that the effect is not merely a statistical one, but has a kind of *necessity* which is characteristic of mathematical discoveries, though a typical learner cannot prove such a thing by using logic. If objects are not only rigid but also impenetrable many other examples of structural causation can be discovered: for example if two centrally pivoted rigid and impenetrable adjacent gear wheels have their teeth meshed and one moves clockwise, the other must move counter-clockwise.

These are simple illustrations of a very wide range of truths about geometry and topology that can be discovered by reflection on interactions between proto-affordances, even though the situations have never been perceived and tested. Sauvy and Suavy (1974) present examples of topological discoveries that can be made by children, and presumably also by suitably designed playful and reflective robots, playing with various spatial structures, strings, pins, buttons, elastic bands, pencil and paper, etc.

Meta-cognitive reflection on invariant features of what is perceived, seems to lie behind the philosophical claim of Immanuel Kant 1781 that mathematical knowledge is both synthetic and non-empirical. This is discussed further in Sloman (1971; 1978, ch 8; 2008b), and in an online presentation on how robots might learn mathematics as a result of reflecting on things they learn about actions and processes in the environment http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#math-robot.

Some of these discoveries are primarily about properties of static structures, such as that angles of a triangle must add up to a straight line. But as a child learns to count, and goes through many counting games and experiments, she may be able to notice recurring patterns and come to realise that they too are not merely statistical correlations but *necessary* consequences of features of the processes. For example, if a set of objects is counted in one order the result of counting *must* be the same for any other order of counting (subject to the normal conditions of counting).

Developing a more detailed analysis of architectural and representational requirements for systems capable of making such discoveries is research in progress. But it is clear that all the cases depend on the fact that an individual can first learn to do something (e.g. produce or perceive a type of process) and then later notice that the process has some inevitable features – inevitable in the sense that if certain core features of the process are preserved, altering other features, e.g. the location, altitude, temperature, colours, materials, etc. of the process *cannot* affect the result.

This also suggests that a Kantian *structure-based* notion of causation is usable alongside the Humean *correlation-based* notion of causation (often expressed in Bayesian nets nowadays). I suspect some other animals, e.g. some nest-building birds and hunting mammals develop Kantian causal reasoning abilities.[2] Similarly, recognition of invariant patterns in sets of sentences leads to logical discoveries made centuries ago by Aristotle and then later extended by Boole, Frege, etc. regarding which patterns of inference are

[2] As discussed here:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/wonac/

valid in virtue of their logical form alone. In other words, whereas Bertrand Russell and some other philosophers tried to reduce all mathematical knowledge to logical knowledge (thought of as a collection of tautologies), I claim that logical knowledge comes from use of meta-cognitive mechanisms involved in discovering mathematical knowledge.

## Reflecting on epistemic affordances

Gibson's notion of affordances for actions that produce physical results can be generalised: In any situation, a perceiver has access to some information about the environment but not all. I.e. that situation has certain "epistemic affordances", defined by what information the individual is capable of acquiring, e.g. by attending to it. While performing many actions, such as moving around and manipulating objects, a child, animal or robot may discover that actions have not only physical consequences, i.e. consequences that satisfy or fail to satisfy physical goals, such as grasping something, or moving to a desired location, but also have the effect of *changing the epistemic affordances. Action affordances* are the possibilities for and constraints on possible actions that can be performed, whereas positive and negative *epistemic affordances* in a situation are the varieties of information available to or hidden from the perceiver.

Moving towards an open doorway gives you access to more information about what is beyond the door. Moving in a plane parallel to the plane of the door frame, changes what information is accessible about the contents of the room: some information will become accessible and some inaccessible. As you move round a house you discover things about the external walls, doors and windows of the house, including their sequential order. You can then use that information to work out the epistemic affordances available by going round in the opposite direction (as Kant noticed) – an essentially mathematical competence at work in a familiar non-mathematical context.

It seems that in the first few years of life a child acquires not only hundreds of facts about actions that alter action affordances, but also hundreds or possibly thousands of facts about actions that alter epistemic affordances. There are many more of these since every slight movement forward, backward, turning, looking up, looking down, moving an object will immediately alter the information available in the environment. At that stage the child will not know these things are being learnt: the meta-semantic competence to reflect on what is going on has not yet developed: how it develops, and what changes occur in forms of representation, mechanisms or architectures are interesting research questions that will have to be addressed before we can design human-like robots. They may also have profound importance for education, especially as children with disabilities (including blindness or deafness) produced genetically or through illness or injury can reach similar end states via different routes, and that may be true also of future robots.

## Epistemic Affordances and Uncertainty

Our ability to predict changing epistemic affordances often reduces the need to compute with uncertain information expressed as probability distributions, reducing the need for probability-based techniques in perception, prediction and planning. Often, an agent that notices that there is some uncertainty about a matter of importance, e.g. because of noise or imprecise sensors, can avoid reasoning with probabilities, by detecting an action affordance that provides new epistemic affordances, reducing or removing the uncertainty, so that simple reasoning or planning suffices.

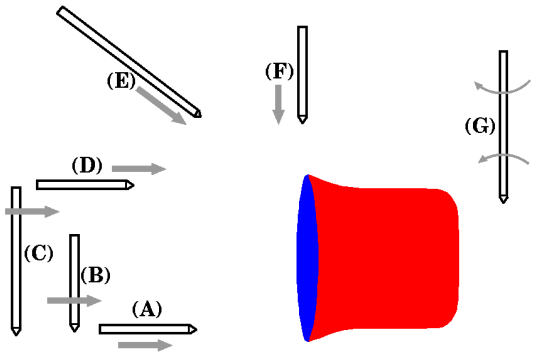


Figure 5: *Various possible configurations involving a pencil and a mug on its side, with possible translations or rotations of the pencil indicated by arrows. If pencil A moves horizontally to the right, will it enter the mug? What vertical change of location of the pencil could make the answer clearer – improving the epistemic affordance? If pencil G is rotated in the vertical plane about its top end will it hit the mug? What translations will make the answer clearer? Similar questions arise for other starting positions of the pencil.*

In Fig 5, the pencils lie in the vertical plane through the axis of the mug. Consider various starting positions and possible translations or rotations of the pencil, and ask: "Will it enter the mug?" "Will it hit the side of the mug?" "Will it touch the rim of the mug?" In some cases there are good epistemic affordances: the answer is certainly "Yes" or certainly "No". Between those situations are "phase boundaries", where epistemic affordances are reduced because the information available is uncertain. But a meta-management system can sometimes tell that there is an affordance for an action that will generate good epistemic affordances, because the pencil can be moved away from the phase boundary to a configuration without uncertainty.[3]

In many everyday situations, we take actions to alter epistemic affordances so as to remove uncertainty. A thirsty individual may see that a mug on the table is within reach, without knowing whether it contains water. Instead of reasoning with probabilities, he may detect an action possibility that will provide new epistemic affordances: standing up would allow him to see into the mug, as would grasping the mug and bringing it nearer.

There is often second-order epistemic information available, namely information that if certain actions are performed some new information will become available and

[3]More examples are in this online discussion paper http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0702

other information may possibly become inaccessible. An animal trying to manipulate some food with a stick poked through a hole may notice the need to check whether the stick will push the food the wrong way. The current situation may not provide that information, but the animal can tell that letting go of the stick and moving round to the side, will make the information available. So a second-order epistemic affordance concerns actions that can be performed that will alter the first order epistemic affordances, i.e. make some things become invisible and other things become visible.

In many situations, the need to use probability distributions can be avoided by using the meta-knowledge that there are "regions of certainty" (ROCs: definitely-Yes and definitely-No regions) with a phase transition representing a fuzzy boundary between the ROCs a "region of uncertainty" (ROU). Using epistemic affordances makes it possible to move out of the ROU into a ROC, e.g. by changing direction of gaze, changing viewpoint, rotating an object, altering direction of movement, changing size of grip, moving something out of the way, etc. etc.

If you don't know whether your current direction of movement will make you collide with the right door frame you may know that if you aim further to the right you will definitely collide with it and if you aim a bit further to the left you definitely will not collide with it, so by aiming a bit further to the left (a fuzzy amount) you can avoid having to reason with probabilities.

## Conclusion

This essay is an example of requirements analysis, paying close attention to ways in which the environment an animal or robot interacts with influences design requirements. In particular, the complexity of the environment and the agents in it dictates a need for various kinds of hierarchical control (continuous and discrete), and various kinds of representation and reasoning, including representations of spatial structures and processes, meta-semantic representations, and representations of proto-affordances and epistemic affordances, which can all be used in reasoning.

By extending Gibson's work, I have tried to indicate both that there are design ideas about meta-cognition that have not yet been explored, except in very simple situations, and to indicate that further research on this may contribute significantly to making machines more human-like. It may also enable us to understand humans better. In particular, examining closely the things that are learnt by a young child concerning both action affordances and epistemic affordances may help us understand the nature of human mathematical capability, and could perhaps lead dedicated teachers to support mathematical learning much better.

Understanding "normal" human and animal forms of perception, learning and development in more detail, may give us deeper insight into brain disorders that disrupt them, and also help us build more intelligent machines. There is far more work still to be done, but no reason to think it cannot be done – provided that we can understand the architectural and representational requirements, and the myriad positive and negative action affordances and epistemic affordances in our 3-D environments.

## Acknowledgements

I have been working on some of these ideas over many years and have learnt from many authors, including colleagues and friends. In the last few years progress has accelerated as a result of discussions with Jackie Chappell about non-human organisms and nature-nurture tradeoffs, and members of the EU-Funded CoSy project, which is developing an architecture framework and toolkit that should be able, when fully developed, to support many of the architectural ideas discussed here (Hawes et al. 2007; Hawes, Zillich, and Wyatt 2007).

## References

Baars, B. J. 1988. *A cognitive Theory of Consciousness*. Cambridge, UK: Cambridge University Press.

Beaudoin, L. 1994. *Goal processing in autonomous agents*. Ph.D. Dissertation, School of Computer Science, The University of Birmingham, Birmingham, UK. http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#38

Brooks, R. 1986. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation* RA-2:14–23. 1.

Chappell, J., and Sloman, A. 2007. Natural and artificial meta-configured altricial information-processing systems. *International Journal of Unconventional Computing* 3(3):211–239. http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609

Cox, M. T., and Raja, A. 2007. Metareasoning: A manifesto. BBN Technical Memo TM-2028, BBN Technologies, Cambridge, MA. http://www.mcox.org/Metareasoning/Manifesto/manifesto.pdf

Cox, M. T. 2005. Metacognition in computation: A selected research review. *Artificial Intelligence* 169(2):104–141. http://mcox.org/Papers/CoxAIJ-resubmit.pdf

Gibson, J. J. 1979. *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Hawes, N.; Sloman, A.; Wyatt, J.; Zillich, M.; Jacobsson, H.; Kruijff, G.-J.; Brenner, M.; Berginc, G.; and Skocaj, D. 2007. Towards an Integrated Robot with Multiple Cognitive Functions. In *Proceedings AAAI '07*, 1548–1553. AAAI Press. http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0701

Hawes, N.; Zillich, M.; and Wyatt, J. 2007. BALT & CAST: Middleware for Cognitive Robotics. In *IEEE RO-MAN '07*, 998–1003. http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0708

Johnson-Laird, P. 1988. *The Computer and the Mind: An Introduction to Cognitive Science*. London: Fontana Press. (Second edn. 1993).

Kant, I. 1781. *Critique of Pure Reason*. London: Macmillan. Translated (1929) by Norman Kemp Smith.

Kennedy, C. M., and Sloman, A. 2003. Autonomous recovery from hostile code insertion using distributed reflection. *Journal of Cognitive Systems Research* 4(2):89–117.

Langley, P., and Laird, J. 2006. Cognitive architectures: Research issues and challenges. Technical report, Institute for the Study of Learning and Expertise, Palo Alto, CA. http://cll.stanford.edu/ langley/papers/final.arch.pdf

Marr, D. 1982. *Vision*. San Francisco: Freeman.

McCarthy, J. 1995. Making robots conscious of their mental states. In *AAAI Spring Symposium on Representing Mental States and Mechanisms*. Palo Alto, CA: AAAI. http://www-formal.stanford.edu/jmc/consciousness.html

Minsky, M. L. 1963. Steps towards artificial intelligence. In Feigenbaum, E., and Feldman, J., eds., *Computers and Thought*. New York: McGraw-Hill. 406–450.

Minsky, M. L. 1968. Matter Mind and Models. In Minsky, M. L., ed., *Semantic Information Processing*. Cambridge, MA,: MIT Press,.

Minsky, M. L. 1987. *The Society of Mind*. London: William Heinemann Ltd.

Minsky, M. L. 2006. *The Emotion Machine*. New York: Pantheon.

Sauvy, J., and Suavy, S. 1974. *The Child's Discovery of Space: From hopscotch to mazes – an introduction to intuitive topology*. Harmondsworth: Penguin Education. Translated from the French by Pam Wells.

Schwitzgebel, E. 2007. No unchallengeable epistemic authority, of any sort, regarding our own conscious experience - Contra Dennett? *Phenomenology and the Cognitive Sciences* 6:107–112. doi:10.1007/s11097-006-9034-y.

Shanahan, M. 2006. A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition* 15:157–176.

Sloman, A., and Chappell, J. 2005. The Altricial-Precocial Spectrum for Robots. In *Proceedings IJCAI'05*, 1187–1192. Edinburgh: IJCAI. http://www.cs.bham.ac.uk/research/cogaff/05.html#200502

Sloman, A., and Chrisley, R. 2003. Virtual machines and consciousness. *Journal of Consciousness Studies* 10(4-5):113–172.

Sloman, A.; Chrisley, R.; and Scheutz, M. 2005. The architectural basis of affective states and processes. In Arbib, M., and Fellous, J.-M., eds., *Who Needs Emotions?: The Brain Meets the Robot*. New York: Oxford University Press. 203–244. http://www.cs.bham.ac.uk/research/cogaff/03.html#200305

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

Sloman, A. 1978. *The Computer Revolution in Philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press). http://www.cs.bham.ac.uk/research/cogaff/crp/

Sloman, A. 1992. How to Dispose of the Free-Will Issue. *AISB Quarterly* 82,:31–32. http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#8

Sloman, A. 2001. Beyond shallow models of emotion. *Cognitive Processing: International Quarterly of Cognitive Science* 2(1):177–198.

Sloman, A. 2002. Architecture-based conceptions of mind. In *In the Scope of Logic, Methodology, and Philosophy of Science (Vol II)*, Synthese Library Vol. 316. Dordrecht: Kluwer. 403–427. http://www.cs.bham.ac.uk/research/projects/cogaff/00-02.html#57

Sloman, A. 2007a. Diversity of Developmental Trajectories in Natural and Artificial Intelligence. In Morrison, C. T., and Oates, T. T., eds., *Computational Approaches to Representation Change during Learning and Development. AAAI Fall Symposium 2007, Technical Report FS-07-03*, 70–79. Menlo Park, CA: AAAI Press. http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0704

Sloman, A. 2007b. Why Some Machines May Need Qualia and How They Can Have Them: Including a Demanding New Turing Test for Robot Philosophers. In Chella, A., and Manzotti, R., eds., *AI and Consciousness: Theoretical Foundations and Current Approaches AAAI Fall Symposium 2007, Technical Report FS-07-01*. Menlo Park, CA: AAAI Press.

Sloman, A. 2008a. Architectural and representational requirements for seeing processes and affordances. Research paper, for Workshop Proceedings COSY-TR-0801, School of Computer Science, University of Birmingham, UK. http://www.cs.bham.ac.uk/research/projects/cosy/papers#tr0801

Sloman, A. 2008b. Kantian Philosophy of Mathematics and Young Robots. Research paper, Submitted COSY-TR-0802, School of Computer Science, University of Birmingham, UK. http://www.cs.bham.ac.uk/research/projects/cosy/papers#tr0802

Sussman, G. 1975. *A computational model of skill acquisition*. American Elsevier.

Wright, I.; Sloman, A.; and Beaudoin, L. 1996. Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology* 3(2):101–126. http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#2

Wright, I. 1977. *Emotional agents*. Ph.D. Dissertation, School of Computer Science, The University of Birmingham. http://www.cs.bham.ac.uk/research/cogaff/