Varieties of Meta-cognition 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

August 3, 2011

Abstract

Some AI researchers aim to make useful machines, including robots. Others aim to understand general principles of information-processing machines with various kinds of intelligence, whether natural or artificial, including 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, 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 philosophical analysis of requirements for robots in complex 3-D environments, surveys varieties of meta-cognition, drawing attention to requirements that drove biological evolution and which are also relevant to ambitious engineering goals.

Contents

*Based on invited Paper for workshop on Metareasoning: Thinking about Thinking, AAAI'08, July 2008. A modified version of this was published in Cox and Raja(2011), This version uses my preferred punctuation and section-numbers, and has a table of contents and other minor differences.

1 Varieties of Requirements and Designs

AI has always included the study of meta-cognition for both scientific and engineering purposes (Minsky, 1968; Cox, 2005). That includes study of various kinds of self-monitoring, self-control and self-discovery, including development of new concepts for self description. My interest in AI started (around 1969) with philosophical and scientific concerns, aiming for designs expressing scientific theories e.g. (Sloman, 1978, Chs 6–10), rather than useful artifacts e.g. (Russell & Wefald, 1991). This study overlaps with philosophy of mind and evolutionary biology: Evolution produced organisms with many different designs, shaped by many different sets of requirements; and we cannot expect to understand all the trade-offs in humans unless we compare alternatives, including non-human animals and possible robots. That involves studying both the space of sets of requirements (*niche* space) and the space of designs that can be compared and assessed against those requirements (*design* space). Such comparisons, instead of using only numerical fitness measures, should, as noted in (Minsky, 1963), include structured descriptions of strengths and weaknesses in various conditions and in relation to various functions, like consumer reports on multi-functional products. A partial analysis is in (Sloman, 2003).

Simply simulating evolution will not yield such comparisons. Another approach, illustrated in (Sloman, 2007a), attempts analytically to retrace steps of biological evolution, especially identifying important discontinuities. Philosophy, especially conceptual analysis, will inevitably be involved in the process. This chapter attempts to identify issues to be addressed in an analytical comparative study. It overlaps with other chapters, but emphasises biological needs and the physical environment.

2 Requirements for organisms and human-like robots

In waking animals, sensors and effectors interact continuously with the environment, and do not need to share a CPU with more central processes. So internal processes, including planning, deciding, self-monitoring, reflecting, and learning, run *concurrently* with sensing and acting, using dedicated machinery, e.g. different parts of brains. This removes the problem of how much CPU time to allocate to meta-reasoning, investigated by many AI researchers, e.g. (Russell & Wefald, 1991), though other constraints can produce similar problems, e.g. if acting

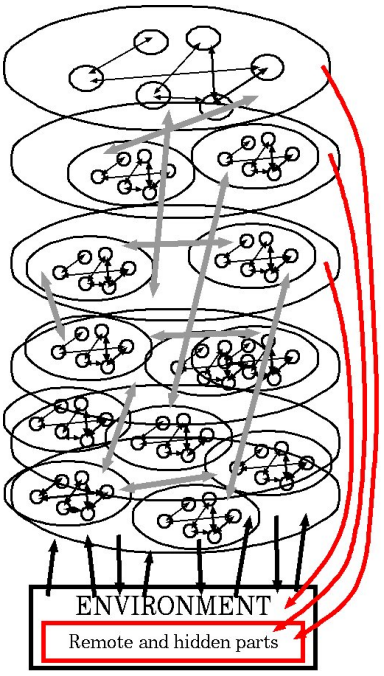

Figure 1: *In animals and robots, concurrently active dynamical sub-systems may vary in many ways, including degree of environmental coupling, speed of change, whether continuous or discrete, what is represented, etc. The longest arrows represent reference from high level subsystems to remote and hidden entities and processes in the environment. Intermediate gray arrows represent information flow between sub-systems. Short black arrows represent sub-state transitions.*

and reasoning about what to do require the agent to be in different locations, or looking in different directions (Sloman, 1978, Ch10). The non-trivial problem of how much dedicated computing power to allocate to each type of function is mostly settled for organisms by evolution.

With dedicated hardware for different tasks, the assumption that intelligent individuals must cycle through "sense→think→act" substates, possibly with meta-reasoning added, can be jettisoned, since architectures include interacting *concurrent* processes of many kinds. (However, some implementations use a single powerful CPU, as argued in (Sloman, 2008b), supporting multiple concurrently active *virtual* machines with different roles.) So arrows in architecture diagrams, such as Figs 1 and 2, unlike flow-charts, can represent flow of information and control between *enduring*, functionally varied, sub-systems, operating at different levels of abstraction, on different time-scales, some changing continuously, others discretely. This has deep implications for forms of representation, algorithms, possible interactions and conflicts between sub-systems, and for trade-offs between design options. Such concurrency was impossible in the early days of AI, as computers had miniscule memories and were far too slow.

The environment (or "ground level") may include arbitrarily complex, partially understood, physical structures and processes, and also other information-users. Intelligent machines, like animals, may start with some "innate" information about the environment, but in many cases will have to develop theories about what sorts of structures and processes can occur in the environment, and how they work. This may involve extending the architecture. Current AI learning mechanisms still lag far behind what animals can achieve.

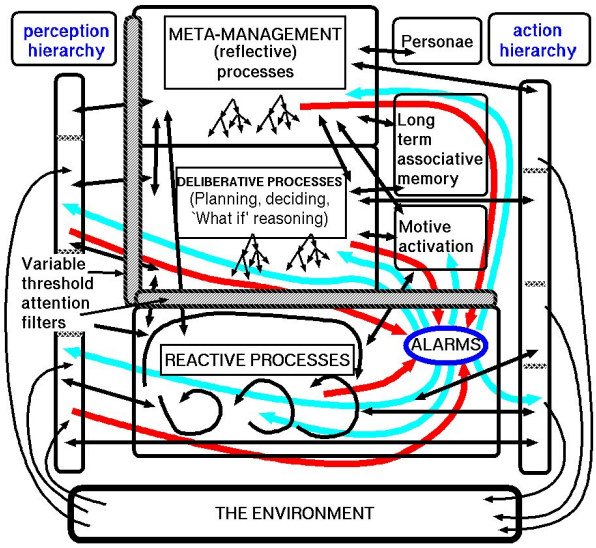

Figure 2: *A sketchy representation of a human-like architecture specification, H-CogAff, developed within the CogAff project (Sloman 2003). Alarm mechanisms mostly monitor passively, but can produce rapid control transfer when needed, sometimes generating emotions. The architecture grows itself while interacting with the environment. This is a special case of the CogAff schema. So far only simpler cases have been implemented. (See Kennedy's chapter.)*

3 Control Hierarchies

Much AI research on meta-reasoning aims to address problems of bounded rationality. However, there is a much older, more general requirement, namely the requirement for hierarchical control. That requirement was "discovered" millions of years ago by evolution and addressed in a wide variety of organisms. Instead of designing a control mechanism that deals with all possible circumstances at a low level of detail, it is often better to provide distinct mechanisms that monitor different things and propose appropriate changes on the basis of what is detected. The changes might modify behaviour immediately, e.g. by changing process parameters or subgoals (e.g. causing gaze redirection), or in the long term by altering sub-modules – as happens in learning and self-debugging systems. Subsumption architectures do the former, using concurrent control at different levels of abstraction (Brooks, 1986). An example of meta-cognition producing long term change was HACKER (Sussman, 1975).

Multiple controllers can sometimes reach conflicting decisions. It is impossible for either evolution or human designers to anticipate all such cases for a complex system functioning in a complex and partly unknown environment. So additional meta-meta-level control subsystems can be useful, monitoring other controllers, and taking action when conflicts are detected. In simple cases, they may modify numerical weights to maximise expected utility (Russell & Wefald, 1991). In more sophisticated designs, dedicated meta-meta-level modules may be able to improve specific modules separately, so as to reduce unwanted interactions, e.g. adding pre-conditions to rules or meta-rules, as in (Sussman, 1975). They may also detect situations

requiring new modules, with their own applicability conditions, and create them by copying and editing portions of older modules, or by using planning mechanisms to create a new complex module for the new context, as in SOAR (Laird, Newell, & Rosenbloom, 1987). For most species this creation of new competences is done only by evolution during phylogeny, though some do it during ontogeny (Chappell & Sloman, 2007).

Meta-level decisions may themselves involve arbitrarily complex problems and the control-systems involved 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). Catriona Kennedy's chapter in this book illustrates "mutual meta-management" by a collection of subsystems each guarding a main system and also other guards. Such systems have not been found in nature, though such a design could be useful in some artificial systems. If telepathy were possible, humans might find mutual meta-management useful!

4 Meta-management and meta-semantic competence

Following (Beaudoin, 1994), we use the label "meta-management" (Fig 2) to emphasise the heterogeneity of "meta-" level functioning, including control as well as monitoring, reasoning, learning, etc. Different meta-management functions can support different types of mental state (Sloman, 2002). Although many researchers regard architectures as unchangeable, in humans, higher level layers develop over several years, including multiple switchable high level control-regimes labelled "Personae" in Fig 2.

Meta-management may use deliberation and reasoning along with reactive mechanisms, e.g. an "alarm" subsystem (Fig. 2) that normally only monitors processes, but can detect situations that need rapid control actions, possibly modifying the behaviour of large numbers of other modules, for instance freezing (in order to avoid detection), fleeing, feeding, fighting or mating. Other options 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, so they will need fast pattern recognition rather than reasoning, and may produce errors. Different effects of such alarm mechanisms in a layered control hierarchy correspond to different types of "emotion" (Sloman, 2001; Wright, Sloman, & Beaudoin, 1996). Different architectures support different affective phenomena (Sloman, Chrisley, & Scheutz, 2005). Acquiring "emotional intelligence" includes learning not to react in some frightening situations, and learning how to modulate "disruptive" control mechanisms to reduce risks, e.g. when controlling a dangerous vehicle. Running alarm mechanisms continuously removes the problem of how often to pause to decide whether to reconsider the situation.

Systems that acquire and use information have *semantic competences*, whether (like neural nets) they use information expressed in scalar parameters or (like symbolic AI systems) they use structural information about states of affairs and processes with more or less complex objects, with changing parts, features, and relationships. In contrast, using information about information, or information about things that acquire, derive, use, contain or express

information, requires *meta-semantic* competences, including the ability to represent things that represent, and what they represent. This includes *representing* having or rejecting (as opposed to merely having or rejecting) beliefs, goals, and plans of both oneself and other individuals.

An individual A with meta-semantic competence may need to represent information I in another individual B where 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 the presuppositions. Further, A may know that a description D1 refers to the same thing as description D2, and therefore substitutes D2 for D1 in various representing contexts. But if B does not know the equivalence, such substitutions may lead A to mistaken conclusions about B's mental states. Dealing with such "referentially opaque" information is more difficult than handling "referentially transparent" forms of representation. Some theorists explore adding new logical operators to standard logic, producing modal belief logics for example. Instead of using *notational* extensions, we can provide *architectural* extensions that allow information to be represented in special "encapsulated" modes, that prevent "normal" uses of the information. Such an encapsulation mechanism can be used for various meta-semantic purposes, such as representing mental states or information contents of other things, counterfactual reasoning and metaphorical reasoning. An example of such a mechanism is the ATT-META system of Barnden (http://www.cs.bham.ac.uk/~jab/ATT-Meta/).

Important research questions include: which animals have which sorts of meta-semantic competence, how and why they evolved, when and how such competences develop in young children, and what brain mechanisms are required to support them. More research is needed on what sorts of meta-semantic competence are required for the meta-management architectural layer in Fig 2, and for the higher level visual capabilities required for seeing someone as happy, sad, puzzled, looking for something, etc., or for intentionally performing communicative actions. Construction of AI models can help us identify requirements and trade-offs, but powerful tools are needed. The SimAgent toolkit (Sloman & Logan, 1999), used in Kennedy's work, was designed to support (among other things) architecture-based meta-semantic competences.

5 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", and "self") is riddled with confusion and muddle. For serious science it is best replaced with a collection of precisely defined labels for special cases, e.g. notions of self-knowledge (McCarthy, 1995). Some self-knowledge based on introspection includes trivial, transient, cases such as a program checking the contents of a register or a sensor reading, and non-trivial cases e.g. architectures with self-observation subsystems running concurrently with others and using a meta-semantic ontology referring to relatively high level (e.g. representational) states, events and processes in the system, expressed in non-transient re-usable

information-structures (Sloman, 2007b).

A system with an architecture allowing introspection to acquire information about its internal states and processes, including intermediate data-structures in perceptual and motor sub-systems, could be said to be self-aware. 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 therefore has the potential for error, notwithstanding arguments 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", is 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.

6 Pre-configured and meta-configured competences

Intelligent systems may start with the ontologies they need for categorising things (as in *precocial* biological species), or, as in some altricial species (Sloman & Chappell, 2005), may develop their own ontologies through exploration and experiment, using mechanisms that evolved to support self-extension, through interaction with a complex, richly structured, changing environment. A distinction can be made between "pre-configured" competences, which are largely genetically determined, and "meta-configured" competences, produced by a succession of acquired competences (Chappell & Sloman, 2007).

Layered development processes can start by learning from the environment how to learn more in that environment, e.g. learning what one can do and what sorts of new information may result – "epistemic affordances" in the environment. Meta-configured learning can include *substantive* ontology development: creation of new concepts not definable in terms of previous concepts. This clearly happens in science (Sloman, 1978, Ch 2), and is also needed in children and intelligent robots. The widely believed theory that all symbols have to be "grounded" in sensory-motor signals is a version of the erroneous philosophical theory of "concept empiricism", as explained in http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#models.

One function of meta-management is discovering the need to modify current theories about the environment, e.g. because predictions have failed. Sometimes abduction can be used to produce a new theory using old concepts, e.g. a theory explaining why a beam of varying thickness does not balance at its midpoint. However some new theories need new concepts referring to the unobserved but hypothesised properties that explain observations, e.g. magnetism. Unfortunately, the search space for abduction of new theories is explosively expanded if additional undefined symbols can be introduced. So learners may need meta-management capabilities to guide the creation of substantially new concepts.

Ontology development is needed not only for coping with the environment, but also for internal meta-management uses, extending the individual's meta-semantic competences, e.g. noticing how one's experience of a rectangular object changes as one views it from different

directions, or noticing that going without liquid for a long time produces an introspectable state.

Meta-semantic ontology extension may result from self-organising capabilities of self-monitoring mechanisms, e.g. using something like a Kohonen net to develop an ontology for intermediate states in perceptual processing, such as tastes, colour sensations, shape experiences, etc. Such concepts of sensory contents may be in principle uncommunicable to other individuals because the concepts are 'causally indexical', i.e. they implicitly refer to the classification mechanism, as suggested in (Sloman & Chrisley, 2003). This may produce philosophical confusions in some future robots.

The space of theories of meta-cognition is vast and unconstrained, except for specific applications. An unexplored constraint suggested in (Sloman, 2007b), is that the theory should explain how different individuals with *the same* initial architecture can reach *divergent* beliefs on many philosophical problems e.g. about the nature of human consciousness, free will emotional states, etc.

7 Affordances, proto-affordances and mathematical meta-cognition

Many researchers assume that the function of vision is to provide information about geometrical and physical properties and relations of objects in the environment. Gibson (1979) argued that organisms need, instead, information about which actions are available to them in particular situations, and which ones will produce desired results: i.e. perception provides information about positive and negative action affordances for the perceiver. This revolutionary proposal was the first step along a major road, though we still have a long way to go (Sloman, 2009). Perception of affordances related to possible actions depends on more fundamental perception of "proto-affordances", namely possible processes and constraints on processes involving motion of 3-D objects and object fragments, whether or not the processes can be produced by the perceiver, and whether or not they are relevant to the perceiver's goals, e.g., seeing how a branch can move in the breeze and how other branches constrain its motion.

Humans can also *reason* about interactions between proto-affordances of different objects, e.g. working out possible behaviours of a machine made of levers, pulleys, ropes and gear wheels (Sloman, 1971). If one end of a long, straight, rigid object is moved down while the centre is fixed, the other end *must* move up. A learner might discover such facts initially as statistical correlations. Later, reflection on what is understood by "rigidity", namely that some feature of the internal structure of the material prevents change of shape, can lead to the realisation that the effect has a kind of *necessity* which is characteristic of mathematical discoveries. 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.

Many truths about geometry and topology can be discovered by reflection on empirically discovered interactions between proto-affordances. Some of the consequences may be

predictable even in situations never previously encountered. (Sauvy & Sauvy, 1974) present examples of topological discoveries that can be made by children, and, I suggest, future 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 Kant's philosophical claim (Kant, 1781) that mathematical knowledge is both synthetic and non-empirical – discussed further in (Sloman, 1971) (1971; 1978, ch 8; 2008a), and in a presentation on how robots, like children, might learn mathematics be reflecting on things they learn about actions and processes in the environment
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#toddlers.

Some discoveries are primarily about properties of static structures, such as that angles of a triangle must add up to a straight line. But a child learning to count, through counting games and experiments, may notice recurring patterns and 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. The discoveries 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 makes it possible for a Kantian *structure-based* notion of causation to be used alongside Humean (or Bayesian) *correlation-based* notions of causation. For reasons given in http://www.cs.bham.ac.uk/research/projects/cogaff/talks/wonac/, it is possible that some other animals, e.g. some nest-building birds and hunting mammals, also develop Kantian causal reasoning abilities.

Similarly, reflection on invariant patterns in sets of sentences could lead to logical discoveries made centuries ago by Aristotle and then later extended by Boole, Frege, etc. regarding patterns of inference that are valid in virtue of their logical form alone. Bertrand Russell tried to reduce all mathematical knowledge to logical knowledge (thought of as a collection of tautologies). I suggest that logical knowledge, like mathematical knowledge, arises from use of meta-cognitive mechanisms reflecting on empirical discoveries, a process not yet modelled in AI.

8 Reflecting on 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. They are linked because an agent can discover that some physical actions change epistemic affordances. Moving towards an open doorway makes more information available about what is beyond the door, whereas moving

sideways both adds and removes information about contents of the next room. 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.

In the first few years of life children acquire not only hundreds of facts about actions that alter *action affordances*, but also myriad facts about actions that alter *epistemic affordances*. Every slight movement forward, backward, turning, looking down, moving an object, etc. will immediately alter the information available. Infants do 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 questions for future research. This may have profound importance for educational policies, especially as children with disabilities (including congenital blindness, deafness or physical deformity) can reach similar end states via different routes, and that may be true also of future robots.

9 Epistemic Affordances and Uncertainty

In a large, complex, partly inaccessible environment neither animals nor machines can achieve complete or certain information. In AI, psychology and neuroscience it is generally assumed that reasoning about probabilities is required for coping with uncertainty and partial information. But in some cases there are simpler and more powerful alternatives, namely, (a) using information about which actions alter epistemic affordances, and (b) using more abstract ontologies that do not require great precision of measurement or control.

Illustrating (a): an agent who 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 alters epistemic affordances, reducing or removing the uncertainty, so that simple reasoning or planning suffices. Examples are moving some object, or changing viewpoint, in order to see more of a partially hidden object or region of space. Often second-order epistemic information is available, indicating that certain actions can be performed to produce new epistemic affordances.

Illustrating (b): instead of using only geometrical descriptions it often suffices to use topological or functional descriptions, or to shift from sub-categories to super-categories. For example, even if you cannot tell the precise distance between two surfaces you can sometimes see that the gap is too small for a nearby armchair to pass through, and sometimes when you cannot tell whether the thing in the distance is a male or a female, you can be sure it's a person, avoiding the need to handle the disjunction and associated probabilities, provided that the person's sex is irrelevant in that context.

Fig 3 indicates possible configurations of a pencil and a mug, and possible translations or rotations, with uncertain consequences. In some cases there are good epistemic affordances, providing clear "Yes/No" answers. Between those situations are "phase boundaries", where epistemic affordances are reduced. A meta-management system can learn about, or discover by

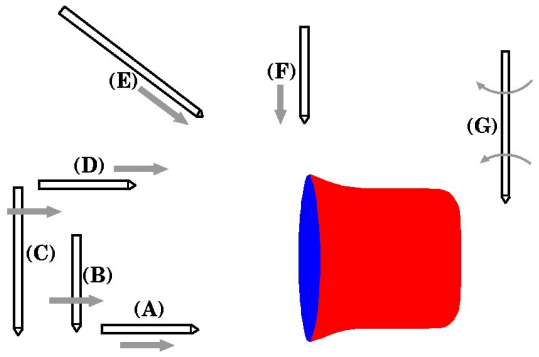

Figure 3: *A mug on its side, with possible locations for a pencil, and possible translations or rotations of the pencil indicated by arrows. If pencil A moves horizontally to the right, will it enter the mug? If pencil G is rotated in the vertical plane about its top end will it hit the mug? In both cases moving the pencil vertically upward removes the uncertainty. The other pencil locations also have associated uncertainty that can be removed by small changes. Different initial moves will extend epistemic affordances for different cases.*

reasoning, that some actions improve epistemic affordances because the configuration moves away from the phase boundary to a region of certainty. A thirsty individual may see that a mug on the table is within reach, without knowing whether it contains liquid. Reasoning with probabilities can be avoided by noticing possible actions that increase epistemic affordances: e.g. standing up to look into the mug, or reaching out to bring it closer. More examples are in http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0702

Sometimes manipulation of 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 fuzzy boundary that is a "region of uncertainty" (ROU). An important type of meta-cognitive learning is discovering when and how it is possible to move from a ROU into a ROC, by performing some action. 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. When you cannot tell whether you are on a course to collide with the right edge of a doorway, you may be able to tell that aiming further to the right will definitely cause a collision and aiming a bit to the left will definitely avoid the collision, without having to reason with probabilities.

10 Conclusion

I have tried to show both how designs produced by evolution, especially designs involving dedicated processors with different functions, escape some of the problems faced by AI researchers considering meta-reasoning in systems based on general computers. But the biological examples produce new problems and opportunities for AI. The CogAff schema presented in (Sloman, 2003), which subsumes Fig 2, provides a framework for exploring, describing and comparing alternative designs with various sorts of meta-cognition, including

varieties that do and do not require meta-semantic competences, a requirement met in humans and other biological organisms, but very few current AI systems.

I have also tried to indicate ways in which detailed studies of the very complex environments in which animals evolved or future robots may have to perform can lead to new requirements and new opportunities for meta-cognition, especially requirements for making use of more varieties of affordance than Gibson identified, including first order and second order epistemic affordances, which can sometimes provide good non-probabilistic ways of dealing with uncertainty.

Gibson's work has been extended, including design ideas about meta-cognition that have not yet been explored, except in very simple situations. Further research on this may contribute significantly to making machines more human-like. It may also enable us to understand humans better.

There is far more still to be done – provided that we can understand the architectural and representational requirements, and the myriad positive and negative action affordances and epistemic affordances in different environments. This may lead not only to more advanced machines, but also to a deeper understanding of what humans and other animals do and how they do it. A better understanding of normal competences could lead to better diagnoses and treatments of genetic or trauma-induced abnormalities. Understanding how young animals learn about first and second order action affordances and epistemic affordances could give us new insights into human mathematical capability, and help dedicated teachers to support mathematical learning more effectively.

Acknowledgements

I have learnt from many people over many years. Recent ideas came from discussions with Jackie Chappell about non-human organisms and nature-nurture trade-offs, and members of the EU-Funded CoSy and CogX robotics projects: http://www.cognitivesystems.org, http://cogx.eu. I thank the editors for their patience and understanding.

References

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.

Beaudoin, L. (1994). *Goal processing in autonomous agents*. Unpublished doctoral dissertation, School of Computer Science, The University of Birmingham, Birmingham, UK.

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

Cox, M. T., & Raja, A. (Eds.). (2011). *Metareasoning: Thinking about thinking*. Cambridge, MA: MIT Press.

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

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)

Laird, J., Newell, A., & Rosenbloom, P. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

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

Minsky, M. L. (1963). Steps towards artificial intelligence. In E. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 406–450). New York: McGraw-Hill.

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

Russell, S. J., & Wefald, E. H. (1991). *Do the Right Thing: Studies in Limited Rationality*. Cambridge, MA: MIT Press.

Sauvy, J., & Sauvy, 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. (1971). Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd ijcai* (pp. 209–226). London: William Kaufmann. (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).

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 P. Gärdenfors, K. Kijania-Placek, & J. Woleński (Eds.), *In the Scope of Logic, Methodology, and Philosophy of Science (Vol II)* (pp. 403–427). Dordrecht: Kluwer. (http://www.cs.bham.ac.uk/research/projects/cogaff/00-02.html#57)

Sloman, A. (2003). *The Cognition and Affect Project: Architectures, Architecture-Schemas, And The New Science of Mind.* (Tech. Rep.). Birmingham, UK: School of Computer Science, University of Birmingham.

(http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200307 (Revised August 2008).)

Sloman, A. (2007a). Diversity of Developmental Trajectories in Natural and Artificial Intelligence. In C. T. Morrison & T. T. Oates (Eds.), *Computational Approaches to Representation Change during Learning and Development. AAAI Fall Symposium 2007, Technical Report FS-07-03* (pp. 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 A. Chella & R. Manzotti (Eds.), *AI and Consciousness: Theoretical Foundations and Current Approaches AAAI Fall Symposium 2007, Technical Report FS-07-01* (pp. 9–16). Menlo Park, CA: AAAI Press. (http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0705)

Sloman, A. (2008a, July). Kantian Philosophy of Mathematics and Young Robots. In S. Autexier, J. Campbell, J. Rubio, V. Sorge, M. Suzuki, & F. Wiedijk (Eds.), *Intelligent Computer Mathematics* (p. 558-573). Berlin/Heidelberg: Springer.

Sloman, A. (2008b). The Well-Designed Young Mathematician. *Artificial Intelligence*, *172*(18), 2015–2034. (http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0807)

Sloman, A. (2009). Architectural and representational requirements for seeing processes and affordances. In D. Heinke & E. Mavritsaki (Eds.), *Computational Modelling in Behavioural Neuroscience: Closing the gap between neurophysiology and behaviour.* (pp. 303–331). London: Psychology Press. (http://www.cs.bham.ac.uk/research/projects/cosy/papers#tr0801)

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

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

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

Sloman, A., & Logan, B. (1999, March). Building cognitively rich agents using the Sim_agent toolkit. *Communications of the Association for Computing Machinery*, *42*(3), 71–77. (http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#49)

Sussman, G. (1975). *A computational model of skill acquisition*. San Francisco, CA: American Elsevier. (http://dspace.mit.edu/handle/1721.1/6894)

Wright, I., Sloman, A., & 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)