Why Some Machines May Need Qualia and How They Can Have Them: Including a Demanding New Turing Test for Robot Philosophers

Aaron Sloman
School of Computer Science The University of Birmingham, UK
http://www.cs.bham.ac.uk/~axs/

Abstract

This paper extends three decades of work arguing that instead of focusing only on (adult) human minds, we should study many kinds of minds, natural and artificial, and try to understand the space containing all of them, by studying what they do, how they do it, and how the natural ones can be emulated in synthetic minds. That requires: (a) understanding sets of requirements that are met by different sorts of minds, i.e. the niches that they occupy, (b) understanding the space of possible designs, and (c) understanding the complex and varied relationships between requirements and designs. Attempts to model or explain any particular phenomenon, such as vision, emotion, learning, language use, or consciousness lead to muddle and confusion unless they are placed in that broader context. in part because current ontologies for specifying and comparing designs are inconsistent and inadequate. A methodology for making progress is summarised and a novel requirement proposed for human-like philosophical robots, namely that a single generic design, in addition to meeting many other more familiar requirements, should be capable of developing different and opposed viewpoints regarding philosophical questions about consciousness, and the so-called hard problem. No designs proposed so far come close.

Could We Be Discussing Bogus Concepts?

Many debates about consciousness appear to be endless because of conceptual confusions preventing clarity as to what the issues are and what does or does not count as progress. This makes it hard to decide what should go into a machine if it is to be described as 'conscious', or as 'having qualia'. Triumphant demonstrations by some AI developers of machines with alleged competences (seeing, having emotions, learning, being autonomous, being conscious, having qualia, etc.) are regarded by others as proving nothing of interest because the systems do not satisfy *their* definitions or their requirements-specifications.[1]

Moreover, alleged demonstrations of programs with philosophically problematic features such as free will, qualia, or phenomenal consciousness, will be dismissed both by those researchers who deny that those phenomena can exist at all, even in humans, and by others who claim that the phenomena are definitionally related to being a product of evolution and, therefore, by definition, no *artificial* working model can be relevant.

Most AI researchers in this area simply ignore all these issues, and assume that the definition *they* use for some key term is the right one (and perhaps cite some authority such as a famous philosopher or psychologist to support that assumption, as if academics in those fields all agreed on definitions). They then proceed to implement something which they believe matches their definition. One result is researchers talking past each other, unawares. In doing so they often re-invent ideas that have been previously discussed at length by others, including theories that were refuted long ago! Boden's new historical survey (2006) should help to reduce such ignorance, but a radical change in education in the field is needed, to ensure that researchers know a lot more about the history of the subject and don't all write as if the history had started a decade or two ago. (Many young AI researchers know only the literature recommended by their supervisors – because they transferred at PhD level from some other discipline and had no time to learn more than the minimum required for completing their thesis.)

Some of the diversity of assumptions regarding what 'consciousness' is and how 'it' should be explained can be revealed by trawling through the archives of the psyche-d discussion forum: http://listserv.uh.edu/archives/psyche-d.html starting in 1993, showing how highly intelligent, and well educated, philosophers and scientists talk past one another. A list of controversies in cognitive systems research on the euCognition web site also helps to indicate the diversity of views in this general area: http://www.eucognition.org/wiki/ Unfortunately many researchers are unaware that their assumptions are controversial.

The rest of this paper discusses two main areas of confusion, namely unclarity of concepts used to specify problems and unclarity of concepts used in describing designs. A new (hard) test for progress in this area is proposed.

Some of the dangers and confusions in claims to have implemented some allegedly key notion of consciousness were pointed out in (Sloman & Chrisley 2003). For example, most people will say, if asked, that being asleep entails being unconscious. Yet many of those people, if asked on another

[1] Everyone who has not yet read the trenchant observations in (McDermott 1981) about claims made by AI researchers should do so now. The arguments apply not only to Symbolic AI, which was dominant at the time it was written, but to all approaches to AI.

occasion whether having a frightening nightmare involves being conscious, will answer 'yes': They believe you cannot be frightened of a lion chasing you without being conscious. Sleepwalking provides another example. It seems to be obviously true (a) that a sleepwalker who gets dressed, opens a shut door and then walks downstairs must have seen the clothes, the door-handle, and the treads on the staircase, (b) that anyone who sees things in the environment must be conscious and (c) that sleepwalkers are, by definition, asleep, and (d) that sleepwalkers are therefore unconscious. The lack of clarity in such concepts also emerges in various debates that seem to be unresolvable, e.g. debates on: Which animals have phenomenal consciousness? At what stage does a human foetus or infant begin to have it? Can you be conscious of something without being conscious that you are conscious of it – if so is there an infinite regress?

The existence of inconsistent or divergent intuitions suggests that the common, intuitive notion of consciousness has so many flaws that it is not fit to be used in formulating scientific questions or engineering goals, since it will never be clear whether the questions have been answered or whether the goals have been achieved. Attempts to avoid this unclarity by introducing new, precise definitions, e.g. distinguishing 'phenomenal' from 'access' consciousness, or talking about 'what it is like to be something' (Nagel 1981) all move within a circle of ill-defined notions, without clearly identifying some unique thing that has to be explained. (As I was finishing this paper the latest issue of *Journal of Consciousness Studies* Vol 14,9-10, 2007 arrived. The editor's introduction makes some of these points.)

Understanding What Evolution Has Done

The inability of researchers to identify a single core concept to focus research on is not surprising, since natural minds (biological control systems), and their varying forms of consciousness, are products of millions of years of evolution in which myriad design options were explored, most of which are still not understood: we know only fragments of what we are, and different researchers (psychologists, neuroscientists, linguists, sociologists, biologists, philosophers, novelists, ...) know different fragments. They are like the proverbial blind men trying to say what an elephant is on the basis of feeling different parts of an elephant.[2]

What we introspect may be as primitive in relation to what is really going on in our minds (our virtual machines, not our brains) as ancient perceptions of earth, air, fire and water were in relation to understanding the physical world. Neither the biological mechanisms that evolved for perceiving the physical environment nor those that evolved for perceiving what is going on in ourselves were designed to serve the purposes of scientific theorising and explaining, but rather to meet the requirements of everyday decision making, online control, and learning, although as the 'everyday' activities become more complex, more varied, and their goals more precise, those activities develop into the activities of science partly by revealing the need to extend our ontologies.

Some will object that introspective beliefs are necessarily true, because you cannot be mistaken about how things seem to you (which is why they are sometimes thought to provide the foundations of all other knowledge). To cut a long story short, the incorrigibility of what you think you know or sense or remember or how things seem to you is essentially a tautology with no consequences, like the tautology that no measuring instrument can give an incorrect reading of what its reading is. The voltmeter can get the voltage wrong but it can't be wrong about what it measure the voltage to be. No great metaphysical truths flow from that triviality.

People who are puzzled about what consciousness is, what mechanisms make it possible, how it evolved, whether machines can have it, etc., can make progress if they replace questions referring to 'it' with a whole battery of questions referring to different capabilities that can occur in animals and machines with different designs. The result need not be some new deep concept corresponding to our pre-scientific notion of consciousness. It is more likely that we shall progress beyond thinking there is *one* important phenomenon to be explained.

What needs to be explained is rarely evident at the start of a scientific investigation: it becomes clear only in the process of developing new concepts and explanatory theories, and developing new ways to check the implications of proposed theories. We did not know what electromagnetic phenomena were and then find explanatory theories: rather, the development of new theories and techniques led to new knowledge of what those theories were required to explain, as well as the development of new concepts to express both the empirical observations and the explanatory theories, and our growing ability to perform tests to check the predictions of the theories (Cohen 1962). We now know of many more phenomena involving energy that need to be explained by theories of transformation and transmission of energy than were known to Newton. Likewise, new phenomena relating to consciousness also emerge from studies of hypnosis, drugs of various kinds, anaesthetic procedures, brain damage, the developing minds of young children, and studies of cognition in non-human animals. Different sorts of consciousness may be possible in a bacterium, a bee, a boa constrictor, a baboon, a human baby, a baseball fan, brain-damaged humans, and, of course, various kinds of robots.

Instead of one key kind of 'natural' consciousness that needs to be explained, there are very many complete designs each of which resulted from very many evolutionary design choices, and in some cases a combination of evolutionary decisions and developmental options (i.e. epigenesis – see Jablonka and Lamb (2005)). For example, what a human can be aware of soon after birth is not the same as what it can be aware of one, five, ten or fifty years later. Likewise, the consequences of awareness change.

Adopting the Design Stance

Although AI researchers attempting to study consciousness start from different, and often inconsistent, facets of a very complex collection of natural phenomena, they do try to

[2] Read the poem by John Godfrey Saxe here: http://www.wordinfo.info/words/index/info/view_unit/1

adopt the design stance (Dennett 1978), which, in principle can lead to new insights and new clarity. This involves specifying various functional designs for animals and robots and trying to define the states and processes of interest in terms of what sorts of things can happen when instances of such designs are working. Compare: different sorts of deadlock, or different sorts of external attack, can arise in different sorts of computer operating systems.[3] The use of the design stance to clarify the notion of free will is illustrated in (Sloman 1992; Franklin 1995). The task is more complex for notions related to consciousness.

But there are serious obstacles. In order to make progress, we require, but currently lack, a good set of concepts for describing and comparing different sets of requirements and different designs: we need ontologies for requirements and designs and for describing relations between requirements and designs when both are complex. Without such a conceptual framework we cannot expect to cope with the complex variety of biological designs and the even larger, because less constrained, space of possible artificial designs. Unfortunately, as shown below, different terms are used by different researchers to describe architectures, capabilities, and mechanisms, and often the same word is used with different interpretations.

Don't All Running Programs Introspect?

McCarthy (1995) and Sloman (1978, ch 6) present reasons why various kinds of self-knowledge could be useful in a robot, but specifying a working design is another matter. Is there a clear distinction between systems with and without self-knowledge? The informal notion of self-awareness or self-consciousness is based on a product of evolution, namely the ability to introspect, which obviously exists in adult humans, and may exist in infants and in some other animals. How it develops in humans is not clear.

Normal adult humans can notice and reflect on some of the contents of their own minds, for instance when they answer questions during an oculist's examination, or when they report that they are bored, or hungry, or unable to tell the difference between two coloured patches, or that they did not realise they were angry. Some consciousness researchers attempt to focus only on verbal reports or other explicit behaviours indicating the contents of consciousness, but hardly anyone nowadays thinks the label "consciousness" refers to such behaviours. Many (though not all) would agree that what you are conscious of when looking at swirling rapids or trees waving in the breeze cannot be *fully* reported in available verbal or non-verbal behaviours. Available motor channels do not have sufficient bandwidth for that task. So most researchers have to fall back, whether explicitly or unwittingly, on results of their own introspection to identify what they are talking about.

We designers do not have that limitation, since we can derive theories about unobservable processes going on inside complex virtual machines from the way they have been designed. The design stance naturally leads to specifications that refer to internal mechanisms, states and processes (in virtual machines[4]) that are not necessarily identifiable on the basis of externally observable behaviours.

From the design standpoint, what 'introspect' means has to be specified in the context of a general ontology for describing architectures for organisms and robots: something we lack at present. Many simple designs can be described as having simple forms of introspection, including systems with feedback control loops such as those presented in (Braitenberg 1984). Many simple control mechanisms compare signals and expectations and modify actions on the basis of that comparison. If learning is included, more permanent modifications result. Those mechanisms all include primitive sorts of introspection. AI problem-solvers, planners, and theorem-provers need to be able to tell whether they have reached a goal state, and if not what possible internal actions are relevant to the current incomplete solution so that one or more of them can be selected to expand the search for a complete solution. Pattern driven rule-systems need information about which rules are applicable at any time and which bindings are possible for the variables in the rule-patterns. Even a simple conditional test in a program which checks whether the values in two registers are the same could be said to use introspection. And inputs to synapses in neural nets provide information about the states of other neurons.

So *any* such system that goes beyond performing a rigidly pre-ordained sequence of actions must use introspection, and to that extent is self-conscious. That would make all non-trivial computer programs and all biological organisms self-conscious.

Clearly that is not what most designers mean by 'introspection' and 'self-conscious'. Why not? The examples given use only *transient* self-information. After a decision has been reached or a selection made the information used is no longer available. Enduring, explicit, information is required if comparisons are to be made about what happens in the system at different times.

Moreover, the examples all involve very 'low-level' particles of information. For a system to know that it is working on a difficult problem, that its current reasoning processes or perceptual states are very different from past examples, or that it has not come closer to solving its problem, it would need ways of combining lots of detailed information and producing summary 'high-level' descriptions, using a meta-semantic ontology, that can be stored and re-used for different purposes. If it also needs to realise that something new has come up that is potentially more important than the task it is currently engaged in, it will need to be able to do different things concurrently, for instance performing one task while monitoring that process and comparing it with other processes. (Some examples relevant to learning to use numbers were given in chapter 8 of Sloman, 1978).

So non-trivial introspection involves: *An architecture*

[3] I call this a study of logical topography. Several logical geographies may be consistent with one logical topography. See http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html

[4] See http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#inf

with self-observation subsystems running concurrently with others and using a meta-semantic ontology 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.[5] Additional requirements can be added to provide more sophistication, including various forms of learning (e.g. introspective, meta-semantic, ontology extension), and self-control mechanisms described below.

Non-trivial introspection goes beyond what is required for perceiving and acting in the world, and even what is required for formulating and testing theories, making predictions, making plans and executing plans. The latter are often implemented in a collection of reactive and deliberative mechanisms, without any concurrently active introspective mechanisms, in typical AI robots – which do many things but lack human-like self-awareness. An early exception was the HACKER program described in (Sussman 1975). But most of what I have read in recent years about machine consciousness ignores all earlier work and attempts to start from scratch.

Muddled Reactions and Deliberations

It is perhaps not surprising that there is confusion about notions as complex and multi-faceted as 'introspection', 'emotion', 'belief', 'motivation', 'learning', and 'understanding'. Unfortunately there is also confusion over terms used to describe much simpler architectures, meeting simpler requirements.

What 'reactive' means, for example, varies from one researcher to another, some of whom restrict it to stateless architectures. That would make reactive architectures of little interest, since stateless systems are incapable of any learning, changing goals or needs, or other features of even the simplest organisms. Other authors allow 'reactive' to refer to mechanisms that can sense both external and internal states and produce both external and internal changes, but restrict the word to systems that cannot represent possible future or past situations that are not sensed, or sequences of possible actions. Examples include behaviour-based robots, systems running feed-forward neural nets, and Nilsson's teleoreactive goal achievers (Nilsson 1994). The vast majority of biological organisms have only reactive mechanisms in that sense. However, there are still very rich possibilities within that framework, though not everyone appreciates them. I was surprised to read in a recent collection on artificial consciousness that whereas a purely reactive robot can continue moving towards a visible target it would be helpless if an obstacle got in the way – surprised because for many years I have been teaching students how purely reactive robots can go round obstacles.[6]

Reactive systems can even deal with goal conflicts: Proto-deliberative systems (Sloman & Chrisley 2005; Sloman, Chrisley, & Scheutz 2005) are a special subset of reactive systems in which a pattern of sensory states can simultaneously trigger two or more internal or external responses, where some competitive mechanism selects between them – e.g. using winner-takes-all to select between fighting and fleeing in response to a threat. Such things are probably common in insects as well as many vertebrate species.

However, purely reactive systems cannot meet the requirements for 'fully deliberative' systems which have the ability to represent, compare, describe differences between, and choose between sequences of possible actions, or explanatory hypotheses, or predictions – all with variable structures. These require special architectural support for construction, manipulation, analysis and comparison of "hypothetical" representations of varying complexity that are not simply triggered by internal or external sensors and may be selected only after complex comparisons, and then possibly stored for various future uses.[7]

There are many intermediate cases between reactive systems and fully deliberative systems, though it is worth noting that all those mechanisms have to be *implemented* in reactive systems.[8] Unfortunately, the word 'deliberative' is another that has not been used consistently in the research community. For instance, some people use the label for what we called 'proto-deliberative' systems above, which includes simple organisms that select between options activated in a neural net. Lumping proto-deliberative systems together with systems that can search in a space of newly constructed reusable possible plans or hypotheses obscures important differences in requirements and designs. See footnote 7.

Varieties of Perception and Action

Many AI architectural diagrams show a complex cognitive system with a small input box labelled 'perception' or 'sensors' and a small output box labelled 'action' or 'effectors', suggesting that there are simple "peephole" channels for input and output. This ignores the richness and complexity of perception and action capabilities in humans (and probably many other animals) and the variety of links between those capabilities and central capabilities. Anyone who works on reading text or understanding speech will know that several levels of abstraction need to be processed concurrently. Likewise speaking, typing, or performing music requires multiple levels of control of output. Similar comments apply to many forms of perception and action, requiring what I have called "multi-window" designs for both (e.g. Sloman & Chrisley 2003; 2005), in contrast with "peephole" perception and action.

The full implications of this are too complex to be discussed here, but it is worth mentioning that if there are multiple concurrent levels of perceptual processing and

[5] The enduring information can be *about* transient events.

[6] Illustrated in some movies of SimAgent demos here: http://www.cs.bham.ac.uk/research/projects/poplog/figs/simagent

[7] Fully deliberative systems and a collection of intermediate cases are described in a still unpublished online working document http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604 Requirements for a Fully Deliberative Architecture

[8] These and related distinctions are presented and discussed in (Sloman 1978; 1993; Beaudoin 1994; Wright, Sloman, & Beaudoin 1996; Sloman 1997; 2002b; 2002a; Sloman & Chrisley 2005; Sloman, Chrisley, & Scheutz 2005; Minsky 2006)

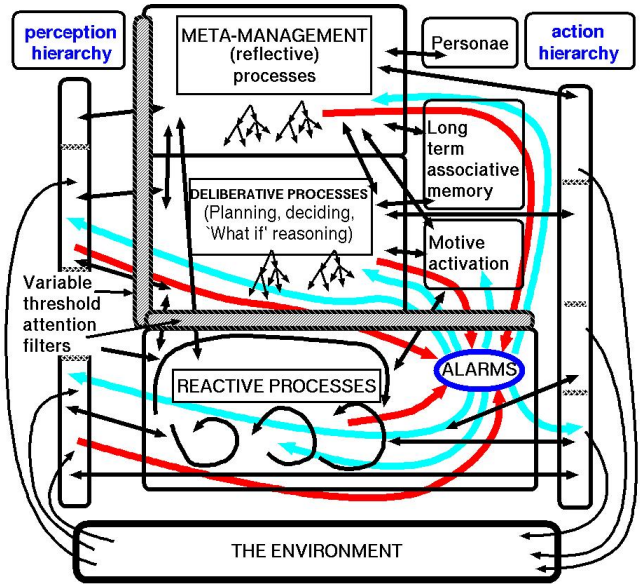

Figure 1: *Sketch of the H-CogAff architecture showing reactive, deliberative and metamanagement layers, with multi-window perception and action, alarms, and personae. Far more arrows are needed than are shown! So far only parts of this have been implemented.*

multiple concurrent levels of control of actions, then that increases the variety of possible contents for self-monitoring. Architectures can vary according to which sub-processes are accessible to introspection. Self-modifying architectures may allow self-monitoring and self-control capabilities to be extended by training (including artistic training, or use of bio-feedback).

For a philosophical robot to start thinking about the phenomenal contents of experience, or 'qualia', its introspective mechanisms would need to be able to access and record the contents of at least some of the perceptual processing subsystems. Different access routes and different internal forms of representation and introspective ontologies would be needed for noticing that you are looking at a river from above, and for noticing the constantly changing details of the swirling rapids. Experiencing the fact that you are seeing a red patch is much less interesting, but also requires introspective access to perceptual sub-processes.

In (Sloman & Chrisley 2003) it was argued that a machine with such capabilities (e.g. implemented using the H-CogAff architecture – see Fig. 1) could use self-organising classification mechanisms to develop ontologies for referring to its own perceptual contents and other internal states. That ontology would be inherently private and incommunicable, because of the role of "causal indexicality" in determining the semantics of the labels used. This explains some of the features of qualia that have led to philosophical puzzles and disputes.

Architectures With Metamanagement

A complex control system can include many reactive and deliberative mechanisms, including feedback control and learning, without having the introspective capabilities described earlier. Systems with the additional self-monitoring capabilities are sometimes described as 'reflective' or 'self-reflective', since some people allow 'reflective' to describe the ability to monitor actions in the environment and learn from mistakes, etc. (e.g. in Minsky 2006).

Monitoring can be a purely passive process, whereas it is often important to *intervene* as a result of monitoring. Various kinds of intervention are possible, including speeding up, slowing down, aborting, suspending, modulating, changing priorities, shifting attention, combining actions, and many more. The label 'metamanagement' (Beaudoin 1994) refers to the combination of introspection and active control based on self-monitoring. A system with metamanagement abilities not only senses and records internal happenings, but can also use that information in controlling or modulating the processes monitored and may even use fully deliberative resources in doing so (see footnote 7).

However it is not possible for *everything* to be monitored and controlled, since that would produce an infinite regress, as discussed in (Minsky 1968), but some subset can be, including the subset discussed in (McCarthy 1995).

What is theoretically possible, and which requirements are met by various possible designs, are open theoretical questions; while what sorts of introspective and self-controlling, i.e. metamanagement, capabilities exist in various biological organisms are open empirical questions. We have identified a need for a special subset of such mechanisms to function as trainable "alarm" mechanisms (Sloman & Logan 1999), closely associated with certain sorts of emotional processes. It is also arguable that metamanagement subsystems may need different monitoring and control regimes for use in different contexts, as humans seem to do: we could call those different "personae", as indicated crudely in Figure 1. Until we know a lot more both about what is theoretically possible and what actually exists, and what the design-tradeoffs are between different possibilities, we can expect discussions of consciousness to remain muddled.

Studying One Thing or Many Things

If there is no unique notion identified by the label 'consciousness', then perhaps in connection with different sorts of organisms and machines the label refers to different complex collections of capabilities.

Suppose we temporarily drop that label and specify every *other* feature of human mentality, including abilities to perceive, remember, notice, forget, focus attention on, shift attention from, reason, plan, execute a plan, reconsider a decision, modify an action, and many affective abilities, such as abilities to have different sorts of desires, inclinations, and preferences, including wanting some states to continue and others to end, and some to begin, and so on. If we can obtain a very detailed set of specifications for everything but consciousness, and use those specifications to produce a working design for a system like a human being in all those respects, including being capable of having all the same dispositional states: not only dispositions to produce externally visible behaviour, but also dispositions

to think, want, like dislike, remember, etc., it is not clear what might be left out that could be added that would make any difference to anything.

Some people think that implies that consciousness (or phenomenal consciousness) is an epiphenomenon: other things can produce and modify it but it has no effects. An alternative is that the notion of consciousness (or qualia) that leads to that conclusion is an incoherent notion – like the notion of the speed at which the whole universe is moving left, without that motion being detectable because all measuring devices that can detect motion are also moving at the same speed in the same direction.

If every other aspect of human mentality can be specified in great detail and emulated in a working system, and if it can be shown what difference different designs occurring in nature or in artifacts make, not just to observable behaviours, but to modes of processing, to energy or other requirements, and to readiness for contingencies that may never occur but would need to be dealt with if they did occur, then all substantive questions about consciousness and other aspects of mind will have been answered, whether philosophers agree or not. (Of course, those driven by fundamentalist religious concerns or a romantic opposition to scientific understanding of human minds cannot be expected to agree: they do not engage in the pursuit of scientific knowledge.)

Factional Disputes

In a more complete discussion it would be of interest to analyse the relationships between different approaches to consciousness and different factions that have arisen in the 50 year history of AI. We can expect to find different designs produced by: those working with symbolic AI systems, using logic or symbolic rules; connectionists using neural nets; researchers on dynamical systems; those working with behaviour based systems; those who have been convinced that physical embodiment is essential; those happy to explore virtual robots interacting with other things in virtual worlds; those dealing only with internet agents, such as trading agents which may become conscious of investments losing value leading to anxiety or fear, and so on.

All these factions seem to me to suffer from a narrowness of vision arising out of conceptual confusions. For example, the current emphasis on the importance of embodiment shifts between tautological triviality (you cannot have perception and action in our physical environment without having a body with sensors and effectors, and the nature of the sensors and effectors will partially determine what an embodied agent can learn about and do in the world), and plain falsehood (an architecture with many of the features of mind that are important for humans cannot be implemented unless it is constantly interacting with the physical and social environment through physical sensors and effectors).

In the past I have explored the idea of a disembodied mathematician concerned only with finding interesting, new, increasingly complex mathematical conjectures, seeking proofs or refutations, trying to improve on old proofs, becoming excited when a problem looks close to being solved, anxious when a proof begins to look flawed, relieved when the flaw is removed, delighted when a very hard problem is finally solved, and so on. Of course, this upsets people from many other factions, but I see nothing inconsistent in the possibility of such a disembodied system. (For a week or two when I was an undergraduate I nearly became such a system while I was spending most of my time lying on my back with my eyes shut trying to prove a theorem I had read about when studying set theory, the Cantor-Bernstein-Schroeder theorem.)[9] A disembodied artificial mathematician of that sort might never experience colours, toothache, the effort in walking uphill, the resistance to pushing a large object, and so on, but it would experience equations, geometric and other structures, a proof being nearly complete, and so on. On completing a proof, or finding a flaw in a previously completed proof, it might have all the non-physical states and processes (including dispositional states in its virtual machine) that are found in the joy or irritation of a human mathematician. Of course, without a body it will not have any feelings in its stomach, tingling of its skin, inclinations to jump for joy. But those are unnecessary for the emotions associated with doing mathematics. (At least they were not necessary for my experiences. You may be different.)

Would such a mathematician have consciousness, emotions, goals, or beliefs? We can avoid futile and interminable debates based on muddled concepts by adopting the design stance and specifying types of consciousness that are available to a disembodied system with a suitably rich virtual machine architecture: e.g. this design is capable of having consciousness of types C_{88} and C_{93} and emotions of types E_{22} and E_{33}. Such proposals will be countered by dogmatic assertions that without full embodiment the mathematician will not have *real* desires, plans, beliefs, consciousness, emotions, etc. Compare denying that a circle is a "real" ellipse.

Shifting the Terms of the Dispute

We can shift the debate about requirements for consciousness in a fruitful way by focusing on phenomena that everyone must agree do exist. For example all disputants must agree that there are people from various cultures who, possibly for multiple and diverse reasons, are convinced that there is something to be discussed and explained, variously labelled 'phenomenal consciousness', 'qualia', 'raw feels', 'what it is like to be something', etc. though they may disagree on some details, such as whether these are epiphenomenal (i.e. incapable of being causes), whether their nature can be described in a public language, whether they can exist in non-biological machines, whether they have biological functions, whether other animals have them, how they evolved, whether it is possible to know whether anyone other than yourself has them, etc. Likewise everyone who enters into debates about the truth of such convictions must agree that there are others who strongly disagree with those opinions.

These disputes involving highly intelligent people on both sides clearly exist, and people on both sides acknowledge

[9] Described in http://en.wikipedia.org/wiki/Cantor-Bernstein-Schroeder_theorem

their existence by taking part in the disputes. So that is something that needs to be explained. Even people who dispute the need for a scientific explanation of qualia (e.g. because they claim the concept is incoherent) must agree on the need to explain the existence of disputes about qualia. So people on both sides of such disputes must agree that an adequate implementable theory of how typical (adult) human minds work should explain the possibility of views being held on both (or all) sides of such disputes.

A New "Turing Test": for a Robot Philosopher

The ability of *one* design to produce robots that favour one or other side in such a philosophical dispute about consciousness should not arise from addition of some otherwise unnecessary feature to the design: it should arise out of design features that have biological or engineering advantages (at least for some species of animal or machine) *independently* of modelling or explaining these philosophical tendencies. Moreover, the same design features, presumably common to all human minds, should explain the possibility of an intelligent robot becoming a supporter of any of the views encountered in disputes about consciousness.

To produce a design suited to this test we need to start by considering only functionally useful architectural requirements for the design of an animal or machine with a wide range of information-processing capabilities, such as humans have, all of which are capable of producing some useful effects, which might help to explain how they evolved. This could include having an architecture that provides metamanagement mechanisms for *internal* self-monitoring and self control, as already described. The detailed specification can be left as a task for designers wishing to show that their robot can pass the robot philosopher test.

To pass the test such a design should enable a robot to notice facts about itself that are naturally described in ways that we find in philosophers who wish to talk about qualia, phenomenal consciousness, raw feels, etc. The very same basic design must also explain why such a robot after studying philosophy, or physics or psychology should also be capable of becoming convinced that talk about qualia, etc. is misguided nonsense. E.g. we should be able to use the same design to model both people like Thomas Nagel, or David Chalmers (1996), and people like Daniel Dennett or Gilbert Ryle (1949). Perhaps such a robot should be capable of reaching the conclusions presented in this paper and proposing this robot turing test.

The functioning model would show how individuals starting with the same sort of genetic makeup can develop in different ways as regards their standards of meaningfulness, or their standards of evidence for theories. Or more subtly, they may develop different ontologies for describing the same portion of reality (as humans often do). In such a situation we may be able to explain what is correct and what is incorrect about the assertions made on both sides, for instance, if the contradictions in their descriptions of the same phenomena arise out of incomplete understanding of what is going on. Ideally we should be able to provide a deep new theory that incorporates what is correct in both sides and exposes the errors made by both sides.

Generalising Bifurcation Requirements

Perhaps a theory of this sort could deal in the same way not merely with disputes about consciousness, but also disputes about free-will, about the nature of affective states and processes, about the existence of 'a self', and about the nature of causation. The design should allow some robot philosophers to become convinced that physical embodiment is essential for mentality, while others argue that purely disembodied intelligences are perfectly possible (as long a physical machines are available on which to implement the required virtual machines). Some should reach Hume's views about causation being nothing more than constant conjunction, while others end up agreeing with Kant that there is something more.

Producing a theory about a design that allows for various bifurcations regarding a host of philosophical problems will require us to answer many questions about how normal, adult, human minds work. It is likely that any such theory will also provide a basis for modelling novel kinds of minds by modifying some of the requirements and showing which designs would then suffice, or by showing how various kinds of damage or genetic malfunction could produce known kinds of human abnormality, and perhaps predict the possibility of types of minds and types of abnormality in human minds that are not yet known.

This work has already begun. As reported above, in (Sloman and Chrisley 2003) a partial specification was given for a machine whose normal functioning could lead it to discover within itself something like what philosophers have called 'qualia' as a result of developing an ontology for describing its sensory contents. Further development of the design may help to resolve questions that currently hinder progress in both AI and philosophy.

The design and implementation of such machines, and analyses of their tradeoffs, could help to unify philosophy, psychology, psychiatry, neuroscience, studies of animal cognition, and of course AI and robotics.

Finally: Major Omissions

There are many things that have not been mentioned here or which require far more detailed discussion. I have assumed that the architectures under discussion will include many affective states and processes, arising from mechanisms for generating new goals, preferences, ideals, likes, dislikes, etc., mechanisms for resolving conflicts, and mechanisms for learning new ways of doing all those things. The previously mentioned "alarm" mechanisms are a special case. These topics have been discussed in more detail elsewhere, especially Simon's seminal (1967) and also (Sloman 1978; 1993; Beaudoin 1994; Wright, Sloman, & Beaudoin 1996; Sloman 1997; 2002b; 2002a; Sloman & Chrisley 2005; Sloman, Chrisley, & Scheutz 2005; Minsky 2006), and in the writings of many other authors. However much work remains to be done on requirements for motivational and related mechanisms. Without that the goals of this discussion cannot be achieved. [10]

[10] See also 'Consciousness in a Multi-layered Multi-functional Labyrinthine Mind': Poster for Conference on Perception, Action

Acknowledgements

My early work in this area benefited much from interactions with Margaret Boden, Luc Beaudoin and Ian Wright. Some of this work was done with Brian Logan, Ron Chrisley and Matthias Scheutz in the context of a Leverhulme Grant, 1999-2002, for research on evolvable architectures for human-like minds. Some it was part of the EU-funded CoSy Cognitive Robotics project. Other recent contributors to the ideas presented are Catriona Kennedy, Nick Hawes, Dean Petters, Jackie Chappell and Arnold Trehub. I have also learnt from many others. (That does not mean anyone agrees with any of this.)

References

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

Boden, M. 2006. *Mind As Machine: A history of Cognitive Science (Two volumes)*. Oxford: OUP

Braitenberg, V. 1984. *Vehicles: Experiments in Synthetic Psychology*. Cambridge, Mass: The MIT Press.

Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York, Oxford: OUP

Cohen, L. 1962. *The diversity of meaning*. London: Methuen & Co Ltd.

Dennett, D. C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

Franklin, S. 1995. *Artificial Minds*. Cambridge, MA: Bradford Books, MIT Press.

Jablonka, E., and Lamb, M. J. 2005. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge MA: MIT Press.

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

McDermott, D. 1981. Artificial intelligence meets natural stupidity. In Haugeland, J., ed., *Mind Design*. Cambridge, MA: MIT Press.

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

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

Nagel, T. 1981. What is it like to be a bat. In Hofstadter, D., and D.C.Dennett., eds., *The mind's I: Fantasies and Reflections on Self and Soul*. Penguin Books. 391–403.

Nilsson, N. 1994. Teleo-reactive programs for agent control. *Journal of Artificial Intelligence Research* 1:139–158.

and Consciousness: Bristol UK, July 2007. http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#pac07

Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson.

Simon, H. A. 1967. Motivational and emotional controls of cognition. In Simon, H. A., ed., *Models of Thought*. Newhaven, CT: Yale University Press. 29–38.

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

Sloman, A., and Chrisley, R. L. 2005. More things than are dreamt of in your biology: Information-processing in biologically-inspired robots. *Cognitive Systems Research* 6(2):145–174.

Sloman, A., and Logan, B. 1999. 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.

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?* Oxford, New York: OUP. 203–244. http://www.cs.bham.ac.uk/research/cogaff/03.html#200305.

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. 1993. The mind as a control system. In Hookway, C., and Peterson, D., eds., *Philosophy and the Cognitive Sciences*. Cambridge, UK: Cambridge University Press. 69–110. http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#18.

Sloman, A. 1997. What sort of control system is able to have a personality. In Trappl, R., and Petta, P., eds., *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*. Berlin: Springer (Lecture Notes in AI). 166–208.

Sloman, A. 2002a. Architecture-based conceptions of mind. In *In the Scope of Logic, Methodology, and Philosophy of Science (Vol II)*, 403–427. Dordrecht: Kluwer. (Synthese Library Vol. 316).

Sloman, A. 2002b. How many separately evolved emotional beasties live within us? In Trappl, R.; Petta, P.; and Payr, S., eds., *Emotions in Humans and Artifacts*. Cambridge, MA: MIT Press. 35–114.

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.