ENGINEERING PHILOSOPHY

COMMENTARY ON "AN ALTERNATIVE TO WORKING ON MACHINE CONSCIOUSNESS", by AARON SLOMAN

Catherine Legg

The concept of consciousness has for the longest time been the domain of philosophers. In fact, it is arguably one of our most favorite concepts, and we are reluctant to cede it to biologists, AI theorists or robotics engineers. Thus the Australian philosopher David Chalmers has distinguished the so-called 'easy problem of consciousness' (What mechanisms perform functions such as receiving sensory inputs?) from the 'hard problem' (Why is all that processing accompanied by *sensations,* and why do they *feel th*e way they do?), arguing that the latter is in principle unanswerable in natural science (Chalmers, 1995). However Sloman was originally dual-trained (in mathematics and physics at Cape Town and philosophy at Oxford) and through a long and very productive career in AI, his work has been unusually philosophically informed. This has included using philosophical tools such as making distinctions and pointing out logical inconsistencies, and drawing ideas from key figures in the history of philosophy such as Kant and Hume, whilst trying to learn from their mistakes and avoid reinventing the wheel. But at the same time his work has been richly grounded in empirical research, with particular reference to building systems to test ideas in practice.

Ever since the wholesale discovery of the experimental method and the ensuing runaway success of the natural sciences in the 17th century, philosophy as a research field has been troubled by the question of its proper role and methodology. In the 18th century David Hume, who was deeply impressed by Newton, influentially divided all knowledge into *relations between ideas*, which are determined a priori (just by thinking) and *matters of fact*, which can only be learned through experiencing the world. This apparently obvious and useful distinction seemed to create a dilemma for the philosophical researcher. On the one hand, he could engage in a priori investigation, which in practice meant largely decomposing arguments as finely as possible in order to assess their validity, and where irreducible premises must be relied on, obtaining them from 'intuition'. This activity risks being devoid of new knowledge or, worse, due to the fallibility of intuition, it risks pronouncing deludedly.

On the other hand, the philosopher could record experiences. However such data-gathering not only risks being rather boring, in many sciences arguably philosophers increasingly lack the specialized training to do it appropriately. Yet retreating from empirical engagement would seem to leave a rather anodyne 'armchair' role (broken only by the occasional excitement of "telling off" other philosophers who tried to leave the chair). It's worth noting that very recently an influential trend in *experimental philosophy* has sprung up to challenge the armchair role ((Knobe and Nichols, 2008), though arguably the pioneer of these ideas was Arne Naess). By 'experimental' its adherents mean the use of quantitative research to address philosophical questions, and by this they mean surveying the intuitions of ordinary people concerning matters such as the true extension of the concept 'knowledge'. We will see however that the experimental character of Sloman's research is in interesting ways the exact opposite of this.

There is a methodological option entirely elided by Hume's dichotomy, however, which is role-modelled by the relatively little-studied scientific discipline of engineering. This is an *architectural* investigation whose key goal may be understood as mapping out 'design spaces'. It seeks to achieve a holistic understanding of what *possibilities* such spaces embody and how different parameters might structure them in nomic and highly inter-connected ways. This object of study does not reduce to either relations of ideas or matters of fact in Hume's terms. To see how this works, consider a simple example – mapping out a design-space for an ideal coffee-maker. Such an investigation is not a priori in that in order to make progress one has to actually build some coffee-makers and use them to make coffee, at least at the beginning. But it is not mere empirical data-gathering either, as the task involves charting many *necessary* relations between parameters such as water temperature, fineness of grind, the heat-conducting properties of the components, and much more, on the final drink.

Such a methodology also differs from much contemporary philosophy insofar as it replaces mere analysis of individual concepts or arguments with a much more synthetic perspective. For under the architectural model, as important as any single parameter is the way in which all interact to form a complex whole (and in our case, produce a tasty coffee. Or not.) A further corrollary of this approach worth mentioning is that it will almost certainly produce *new concepts* which could never be anticipated prior to actually designing and building the object(s) of interest. (In the case of coffee-making, one might mention concepts such as 'espresso' and 'crema'). Such concepts can claim the status of discoveries as much as any law of interaction between parameters. Thus Sloman writes:

> ...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 concepts to express both the empirical observations and the explanatory theories...(p. 5)

Few philosophers seem to recognize that there might be a posteriori discoveries concerning meaning. Thus one might argue that the problem with Hume's claim that 'relations of Ideas' is a priori was that he had a non-scientist's cosy assumption that the meaning of ideas must be easily accessible to introspection. A notable exception is the founder of philosophical pragmatism Charles Peirce[1], and Sloman's scientific approach seems to have rendered him a fellow-traveller to these ideas.

So now if we apply all this to consciousness, what is the result? Sloman seeks to show:

> ...how unclear concepts of common-sense (and philosophy) could be replaced by more precise and varied architecture-based concepts better suited to specify what needs to be explained by scientific theories. (p. 4)

This is where he differs with the "experimental philosophers'" claim that the experimental method should require philosophers to chart much more exactly, and rely on, ordinary people's intuitions about traditional philosophical concepts such as consciousness. This is precisely the wrong direction to go for, sadly:

> The supposed common, intuitive notion of consciousness is mythical, and instead there is a family of different notions, with so many flaws that none of the common labels is fit to be used in formulating scientific questions or engineering goals. (p. 3)

In the rest of the paper Sloman addresses a number of these different notions in turn, in each case exposing unclarities in current discussions, and suggesting that apparent yes-no questions in fact hide a spectrum of positions (again, often across multiple parameters), which when uncovered allows for a vastly more nuanced debate. In section 5, he addresses *introspection*. This is frequently taken to be a necessary or even necessary and sufficient condition for consciousness. But what is meant by it exactly, from a design standpoint? Does a system's having any access to information concerning its internal states qualify? How about "...a simple conditional test in a program which checks whether the values in two registers are the same (p. 7)"?

Reactiveness is a quality often taken to mark the absence of consciousness in a system. But again, what does this mean? Architectures with no internal states at all? Or only those that, "cannot represent possible future or past situations that are not sensed… (p. 8)"? Another option is 'proto-deliberative' systems that can make choices about their responses via simple mechanisms. Sloman suggests that studies of the animal kingdom would show a rich palette of options here. At the other end of the spectrum lies 'fully deliberative systems', which can perform:

> ...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 (p. 9)

although the term 'deliberative' is itself in need of further unpacking.

Perception is another frequently oversimplified notion. Sloman laments the 'peep-hole' conception

[1] The author of this review explores this aspect of Peirce's pragmatism in detail in Legg [2005]

taken for granted by many researchers, and distinguishes it from 'multi-window' perception where several levels of abstraction may be processed concurrently. This is not even yet to specifically address the issue of the system's monitoring its own internal states. After this Sloman levels his sights at the concept of *reflectiveness*. This seems to mean some kind of metamanagement whereby a system:

> ...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. (p. 11)

However how to build such a system is an open empirical question, and there is so much we can only learn by actually trying out specific implementations and exploring the design-tradeoffs therein. (For example, what exactly is the role of emotion in metamanagement? Or personae?)

Sloman then turns to *qualia* – the feature which we have seen Chalmers pegs as absolutely essential for and constitutive of consciousness (conveniently inaccessible in principle to scientific study). He boldly proposes to do without them:

> If every other aspect of human mentality can be specified in great detail and emulated in a working system...then all substantive questions about consciousness and other aspects of mind will have been answered, whether philosophers agree or not. (p. 12)

Of course Chalmers is likely to say that this response begs the question against him, but what further question *could* there be? The dialectic raises interesting questions about answering questions in philosophy. Are all answers to questions in propositional form? Or might an architecture answer a question? Again this option seems to suggest a synthetic mode of thought not common in mainstream philosophy today. Sloman's discussion also opens the arguably neglected possibility that 'the qualia issue' might in fact turn out to be a host of issues once relevant design-spaces are charted.

Sloman finishes the paper with an ambitious high-level design-challenge. He notes that different philosophical theories of mind give rise to different notions of what is essential in designing one. A well-worked example concerns embodiment - could there be a disembodied pure mathematician mind? Would we call this a mind? As always, the answer is:

> 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 C88 and C93 and emotions of types E22 and E33...(p. 13)

Pace possible quarrels regarding spurious discreteness of the numbered 'consciousness-kinds', this makes sense. In a thought-provokingly reflexive move, he also points out that we can also try to understand our raising *such questions as these* from the design stance:

> These disputes involving highly intelligent people on both sides clearly exist, and people on both sides acknowledge their existence by taking part in the disputes. So that is something that needs to be

explained. (p. 14)

The Turing Test famously contributed to the debate concerning whether machines could *think* by operationalizing those features of the concept Turing deemed capable, and ignoring the rest. His exact words bear repeating in the current context:

> I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it... (Turing, 1950, p. 433)

This other closely related question was, famously, the engineering problem of whether a computer might pass undetected as such during a 'conversation' of approximately 5 minutes with it and other humans (with all interlocutors hidden from sight). It is widely accepted that Turing's was an invaluable contribution to the debate over artificial intelligence. Sloman suggests a new Turing Test, but he is not so much seeking to operationalize and/or replace consciousness itself as *our ability to argue about it* (an interesting shift which he doesn't fully explain the reason for). Thus he proposes that *philosophers seek a design for a robot philosopher*, noting:

> 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 (pp. 14-15).

He suggests the result will be "...a deep new theory that incorporates what is correct in both sides and exposes the errors made by both sides (p. 15)".

At this point I become a little sceptical. Are we sure that the positions in disputes between minds can in all cases be traced back to *design features* of those minds? What if more arbitrary and accidental factors are at work, for instance where one went to graduate school? (Though perhaps robot philosophers might be more principled.) Nevertheless, Sloman claims that the design work has begun in his own research programme. If you share his vision please join him!

REFERENCES

Chalmers, D. (1995). "Facing Up to the Problem of Consciousness", *Journal of Consciousness Studies* **2 (3)**, pp. 200-219.

Knobe, J. & Nichols, S. (eds.) (2008). *Experimental Philosophy*. Oxford University Press.

Legg, C. (2005). “The Meaning of Meaning-Fallibilism”, *Axiomathes* **15 (2)**, pp. 293-318.

Turing, A. (1950). “Computing Machinery and Intelligence”. *Mind* **49**, pp. 433-460.