Book Review

Mind and Mechanism, by Drew V. McDermott

Peter Carruthers

Department of Philosophy, University of Maryland, Skinner Building, College Park, MD 20742, USA

This book is a defense of physicalist views of the mind and of consciousness. More specifically, it defends the computational theory of mind from an AI perspective. It is clearly written (though not always easy to read), mostly well informed, and generally well argued. It makes a valuable addition to the literature in this area.

McDermott’s first chapter introduces the problem which the very existence of phenomenal consciousness raises for physicalism. (Phenomenal consciousness is the sort of consciousness which is involved when one undergoes experiences with a subjective ‘feel’ or ‘what-it-is-likeness’.) How can the technicolor phenomenology of our inner lives be created out of the physical interactions of neurons in our brains? He explains how many people have been tempted to answer this question negatively, and to opt for some form of ontological dualism instead. He also explains how the resulting position gives rise to insuperable problems of various sorts. He then sketches the computationalist approach to the problem which will be taken in the remainder of the book.

Chapter two is about AI—but not about the idealized, fictional, AI of philosophical thought-experiments. It is about the real thing: sketching the real progress which has been made by AI researchers in recent years in tackling problems in vision, navigation, language processing, and so on. For someone such as myself, with little working knowledge of AI, this was all very useful and informative. It was particularly interesting to see how (both here and throughout later chapters) assumptions of modularity are now ingrained in the practice of virtually all AI researchers.

My own area of research is the philosophy of psychology, where I interact quite a bit with evolutionary, developmental, cognitive, and comparative psychologists. The question of the modularity of mind is much debated in these circles, with a minority arguing in favor (quite persuasively in my view), but a significant majority either fighting against, or resting unthinkingly with the common-sense picture of the mind as some sort of holistic
system (or general-purpose computer). It is interesting to see so clearly from McDermott’s book that the modularists in this debate have AI firmly on their side. Indeed, it looks as if the opponents of modularity must now face a stark choice: either to give up their position, and embrace modularity, or to give up believing that minds are computationally realized. But in the latter case, where would they go then? As McDermott argues in this book, computationalism is the best (perhaps the only) hope we have of understanding how the mind can be realized in physical mechanisms.

Chapters three and four of the book present McDermott’s computational theory of phenomenal consciousness, and reply to a variety of objections to it. This is, obviously, the heart of the book. I shall say more on it below, with some critical comments. But the basic idea is that perceptual states get to be phenomenally conscious by virtue of being modeled within the agent’s self-model. The account thus falls under the umbrella of so-called ‘higher-order representation’ theories of phenomenal consciousness, of which there are a number of different varieties on the market (including one developed by the present reviewer; see my *Phenomenal Consciousness*, CUP 2000). But quite where McDermott’s view falls in this spectrum of possibilities is much less easy to discern, as we shall see—this is one place where greater clarity and explicitness, and a closer acquaintance with the relevant philosophical literature, would have been a help.

Chapter five mounts a spirited defense of both the coherence and explanatory potential of computationalism as a scientific research program. This is against critics who allege that the very idea of ‘symbols in the brain’ makes no sense, as well as those who say that, even if this did make sense, there would be no objective, observer-independent, way of ascribing meanings to those symbols. There is lots of interesting material in this chapter. But very roughly, McDermott’s overall response to both sets of critics is similar. It is to point out that the notions in question (symbols and their meanings or intentional contents) are embedded within viable and progressive scientific research programs (cognitive science and cognitive psychology respectively). So the properties with which they deal have just as much right to be regarded as real as do those postulated by any other science, such as neuron, or neutrino. My view is that this response gets the matter exactly right.

Finally, in Chapter six, McDermott explores the consequences of physicalism for ethics and religious belief. These are issues which aren’t often raised by philosophers of mind in the context of their discussions of physicalism and computationalism, so it is good to see them aired; even if McDermott’s conclusions aren’t wholly convincing.

Having surveyed the scope of the book, let me now focus on its treatment of phenomenal consciousness in particular. As noted above, McDermott advertises himself as defending a higher-order theory of consciousness, maintaining (very roughly) that a phenomenally conscious sensation will be one of which the subject is aware. But in fact, many of the considerations he introduces in developing and defending his theory are purely first-order in character. This may have led him to under-estimate how difficult it would be to build a computational system which is phenomenally conscious, and to over-estimate the likely distribution of phenomenal consciousness within the animal kingdom. Let me elaborate.

When the idea of a ‘self-model’ is first introduced, it is a model of oneself as a physical system which is actually being talked about. Animals and robots need to be able to locate themselves on their mental maps of the environment in relation to a perceived scene. And in predicting the likely movements of perceived objects they have to factor in the effects of
their own intentional movements. But none of this requires any higher-order representation of the agent's own mental states. For example, in order to predict how a moving ball will change trajectory given that I have formed an intention to kick it, I don't have to represent my own intention qua mental state. Rather, my physics module just has to receive as input a representation of the intended movement of my leg in relation to the moving ball. And that can surely be generated from my intention without anything explicitly representing it as an intention.

Similarly, when discussing how phenomenal 'qualia' arise in a system, McDermott talks about how our color judgments, in the end, have to 'bottom out' in similarity-judgments, without us having any access to the underlying basis for those discriminations. Asked to say how I know that a perceptually presented object is blue, all I can really do is say that it is similar to other blue objects. But there is nothing higher-order involved here either. McDermott may have been misled at this point by having the imagined system express the similarity by judging that the object 'looks like' other blue objects; and the use of this phrase may suggest a capacity for higher-order representation of perceptual states as such. But not so. This use of 'looks like' is just a notational variant on 'is similar to (in certain delimited respects, de facto accessible to vision)'. Nothing higher-order is required.

It may be for these reasons that McDermott thinks that phenomenal consciousness will be quite widely distributed in the animal kingdom (extending to cats and dogs as well as apes, he tells us, and presumably therefore to all mammals at least), despite his endorsement of a higher-order account. And it may also be for these reasons that he is tempted to claim the following, as he does on p. 215:

it seems as if any computational entity that dealt with a physical environment that included its own body would have to have a model of itself as a perceiver and decision maker; and in that model the entity and events involving it would have to be labeled as having the features of phenomenal consciousness. [Emphasis added.]

But there is no warrant for these claims. In fact, everything we know from developmental and comparative psychology suggests that higher-order representations aren't easily come by. Human children only begin to represent their own representational states as such, and to distinguish between the way things are and the way they seem or appear, at around the age of four years; yet of course they are agents long before that. And after nearly two decades of investigation of the higher-order mentalizing abilities of our nearest animal cousins, the great apes, the jury is still out on the question whether or not they possess any such abilities. For sure, apes live in complex social groups and are excellent behaviorists, at least, with sophisticated abilities to predict and manipulate the behavior of their fellows. But it is still hotly disputed whether or not they can represent the representational status of internal mental states in themselves or others, and whether or not they can draw an is/seems distinction. Unfortunately, McDermott seems unaware of these literatures.

Although he shows greater awareness of the existing literature on consciousness, McDermott seems not to have thought very hard about how to situate his own positive proposals in relation to that literature. He does, it is true, distance himself from the views of Dennett, as well as the earlier views of the present reviewer, by denying that there is any constitutive connection between consciousness and language—and rightly so, I now
think. But he never clearly says whether he believes that it is higher-order thoughts about our perceptual states which renders them phenomenally conscious, or whether we need instead to have higher-order perceptions of those states. He does often use the language of perception. For example, he talks about the “ability to perceive the output of sensory systems” (p. 106; cf. p. 133). On other occasions he seems to suggest that any sort of higher-order representation of perceptual states would do (p. 126). And on yet others he uses language strongly suggesting that it is a higher-order conceptualized belief which he has in mind. Thus on p. 203 he writes:

> to have a sensation of red is to have an occurrent representation to the effect that this object I seem to see has a shade of the following type (the essence of the type consisting in the similarity relations it bears to various color samples). [Italics in original.]

This lack of clarity is unfortunate. Although it doesn’t entirely denude McDermott’s book of value, by any means, it does have the consequence that his theory is much harder to evaluate, and so is correspondingly much less useful, than could otherwise have been the case.

There are some passages in the book which suggest that McDermott’s view can be distinguished from most others in the literature by not actually requiring the occurrence of any first-order perceptual state for phenomenal consciousness to occur, and by allowing that the mere presence of a higher-order belief would be sufficient. He often emphasizes that beliefs in the self-model are self-fulfilling. And at one point he imagines a case where an entirely spurious belief to the effect that I have a painful burning in my right leg is created, in the absence of the usual underlying perception which such beliefs are about; and goes on to conclude that such a case would be psychologically indistinguishable from the real thing, by virtue of the self-fulfilling nature of the self-model. He concludes, “The difference between a true sensation report and a false sensation report vanishes” (p. 208).

If this is McDermott’s view—that phenomenal consciousness exists whenever there is a higher-order belief to the effect that a perceptual experience is occurring, whether or not any such experience is really present—then it is surely unacceptable. When I undergo a phenomenally conscious experience of red I don’t just find myself judging, blindly, that such an experience is taking place. Rather, if I make such a judgment it will be grounded in a fine-grained awareness of the character of my perceptual state. I am aware of subtle changes and gradients in the nature of that experience, for example, which slip through the mesh of my conceptual net. (Compare: when I make a first-order judgment of ‘red’, this will be similarly grounded in a more fine-grained awareness of the property recognized.) And surely nothing could count as phenomenal consciousness which didn’t involve some such fine-grained awareness. And there is, surely, no reason at all to think that this richer architecture couldn’t be computationally realized. McDermott’s sort of pure higher-order-belief view certainly isn’t forced on us by the demands of computationalism.