Virtually Philosophy

Dan O'Brien
University of Birmingham, UK.

Mind and Mechanism. McDermott, Drew V. MIT Press. Cambridge, Massachusetts, 2001. Pp. viii+ 262.

McDermott aims to show how the mind-body problem can be solved. He targets the "hard problem": how it is that physical entities can have conscious experience. He concludes that we are biological computers. Distinguishing between vehicle and process theories, he adopts the latter: "what's important about a neuron is not the chemicals it secretes or the electrical potentials it generates, but the content of the information encoded in those physical media" (p. 37). He is thus an advocate of AI: if a robot can process information in the way that our brain does then such a robot is also conscious.

In chapters 1 and 2, McDermott gives a clear account of the motivation behind, and problems associated with dualism. He provides a (long) chapter on the history of AI and focuses on two themes. First, the project of AI should not concern itself with searching for a *general* algorithm underlying all thought. Specialisation is the key. He illustrates this in a survey of current research into robot vision, movement, and language. Second, he investigates the distinction between connectionism and GOFAI (good old-fashioned AI): "the bottom line is this: if someone tells you that a system is "symbolic," they have told you almost nothing" (p. 188). It is not just digital computers that manipulate symbols; neural networks have representational abilities, and thus, are also "symbolic". And: "neuron firing patterns are nature's way of making do without a digital computer" (p.188).

In chapter 3 McDermott introduces his theory of consciousness by first considering the windows on one's computer desktop, and then, more controversially, free will. The MS Word window only exists because it is modelled within the computational syntax of the computer. Analogously, "a system has free will if and only if it makes decisions based on causal models in which the symbols denoting itself are marked as exempt from causality" (p. 98). Any strong sense of libertarian freedom is therefore denied, and along with that, the traditional philosophical questions of autonomy and responsibility. Consciousness, too, is a product of the way that a system models itself. To interact in the complex way that we do with our environment our brains must carry information about our place in relation to that environment. We must have a self-model. Robots, too, can have such a model. And, in order to act, a robot (let's call him Huey) must have programmed preferences. These must ultimately rest on Huey's sensitivity to certain stimuli□ such as heat □ that are classified as "intrinsically not-likeable" (p.102). The crucial claim is that such intrinsically classified sensory states play the same role in Huey as sensations of heat play in us. Similarly, Huey can be programmed to distinguish colours, and to do so there must be recognitional states in him that play the role of our colour qualia. So far, then, Huey has certain intrinsic preferences and sensory abilities, and, he behaves in much the same way as we do: he avoids fire, and perhaps prefers green rooms to red rooms. Huey acts as if he is conscious, or in McDermott's phrase, he has "virtual consciousness."

The key claim of the book is that consciousness is just virtual consciousness. We are conscious because the neurophysiology of our brains instantiates a computationally structured self-model. If Huey deals with his environment in a sufficiently complex way then he must be agile (able to manipulate objects), have a model of his own body, be a decision maker, and, as a necessary component of such computational intelligence, we find conscious awareness, free will, and emotion.

McDermott takes an anti-Cartesian position. A conscious self that surveys its internal Cartesian theatre within which qualia take centre stage does not exist. The self-model of a system creates the illusion of such a self in the same way an algorithm creates the MS Word window. McDermott allies himself to the higher order thought theorists: there is no phenomenological, felt quality to experience; you just believe that there is.

In chapter 4, the usual suspects of the philosophy of mind are considered. How can a materialist theory of the mind account for the possibility of zombies, the knowledge acquired by Jackson's Mary, inverted qualia, what it is like to be a bat, and the Chinese room? None of these objections to materialism are found pressing.

Chapter 5 addresses the question of the representational abilities of computational systems. Seen as "syntactic engines" computers turn input into output. The interesting computers, however, are those that can be interpreted as possessing semantic content; those whose relationship with their environment allows that their internal workings can be seen as *about* that environment, as having intentionality. McDermott ascribes to a causal theory of content: the representational content of a neuron-firing pattern is determined by the causal links those neurons have to the environment.

McDermott presents a comprehensive treatment of dualism, a useful survey of AI, and an interesting and radical theory of consciousness. There are, however, certain problems with the book.

There is too much speculation concerning future technological advances. While not claiming that current computers are conscious, he fears that the reader may not accept his position because he can only provide "sketches of arguments until cognitive science has advanced far enough to fill in some details" (p. 210). His conviction that the philosophical ground will shift with (possible) technological advances indicates the philosophical naivety of the book. Whether or not AI is ever *actually* successful, McDermott's theory of consciousness remains a substantive, interesting, and highly controversial thesis concerning the basis of consciousness and the mind.

McDermott gives philosophical problems only superficial treatment. His discussion of the causal account of content ends with: "perhaps intentionality itself can be explained in terms of informational meaning, thus erasing its seeming distinctiveness" (p. 199), with developments in cognitive science left to fill in the explanation. He misunderstands the depth of Quine's (1960, pp. 26-79) radical translation argument: even "if we could measure events in the brain better", it is not clear how that would help narrow down the alternative interpretations of what our thoughts are about (p. 205). Chapter 5 ends with a discussion that ranges over foundational epistemology, the Kantian distinction between noumena and phenomena, and solipsism. McDermott claims that Kant's metaphysics is ruled out because science investigates not mere appearances but "how things really are"; and, solipsism is refuted since, in principle, we can come to identify our own experiences with the theoretical constructs of cognitive science using a "cerebroscope," and then use these constructs to "locate experiences in observed minds" (p.214). The philosophy here is underdeveloped. To the cognitive scientist, little idea is given of the depth of the philosophical problems involved; and to the philosopher, McDermott's responses will be unpersuasive as they stand.

Chapter 6 confirms that the book is philosophically over ambitious. Here McDermott turns to morality. He provides an unconvincing argument against robots having a sense of morality, humour, aesthetics or love: one "can't count an artefact as really loving or laughing if changing a few program variables would radically change what it loved or laughed at" (p. 218). AI and other advances in technology will also raise ethical problems, and to help us deal with these a moral theory is called for. Utilitarianism, Kantian ethics, nihilism, and moral relativism are all sketched. Not persuaded by such theories, McDermott suggests turning to God: "I find the world to be morally incomprehensible without being able to adopt God's view of it, and physically inexplicable unless there is something outside of it that explains why it exists. I realize perfectly well that just saying the word "God" does not solve either problem. What it does is acknowledge the *holiness* surrounding them, and express faith that they are linked" (p. 237). There are some general considerations concerning whether religion is compatible with science, and a very unexpected conclusion. McDermott makes the unsubstantiated claims that God created the laws of physics and left us to evolve; that perhaps a belief in God is an inevitable aspect of intelligence; and that God may be essential for modern democratic civilization to evolve. For all this we should thank him and pray. The bizarreness of all this is tempered by the suggestion that we may possibly be praying to ourselves; that just like free will and consciousness, God too is only virtual.

McDermott worries that "many readers are going to find the central argument preposterous, obscure, or inadequate" (p. 137). It may be some of those things; and thus, it would be prudent to develop the key argument, rather than become sidetracked on such general and unpersuasive philosophical discussion.

Bibliography

Quine, W. V. O. *Word and Object.* MIT Press, Cambridge, Mass, 1960.