Published in the Newsletter of the *Cognitive Science Society*. Vol 22, Issue 2, June, 2002
http://www.cognitivesciencesociety.org/newsletter/June02/index.html

A Connecticut Yalie in King Descartes' Court

A review of ***Mind and Mechanism***

by Drew McDermott,

MIT Press, Cambridge, MA, 2001

Eric Dietrich
Philosophy Dept.
Binghamton University
Binghamton, New York

dietrich@binghamton.edu

Valerie Gray Hardcastle
Science and Technology Studies Program
Virginia Tech
Blacksburg, Virginia

valerie@vt.edu

What is consciousness? Of course, each of us knows, privately, what consciousness is. And we each think, for basically irresistible reasons, that all other conscious humans by and large have experiences like ours. So we conclude that we all know what consciousness is. It's the felt experiences of our lives. But that is not the answer we, as cognitive scientists, seek in asking our question. We all want to know what physical process consciousness is and why it produces this very strange, almost mysterious, phenomenon of felt experience.

Traditionally a philosophical problem, but desultorily picked at by neuroscientists and psychologists, we now have a proposal from a computer scientist, Drew McDermott, for how a computational mechanism could be conscious. And a fine proposal it is. Though his book is short, it contains a lot of interesting ideas. The centerpiece is a computational theory of consciousness. Along the way, we are treated to a refreshingly frank assessment of the current status of AI, accessible descriptions of what AI can do,

an insightful discussion of the nature of representations and their semantics, and some very good discussion about the relationship between being mechanical and yet remaining human -- a topic usually avoided by philosophers, even though they are the ones who should be discussing it. But, are we really done? Is there actually a theory computational theory of consciousness in this book. Is it just a matter now of working out and then implementing the details? Is our future that rosy? Of course, you should get the book and decide for yourself. Here, we provide some comments that might make your decision easier.

McDermott has a straightforward, comfortable style that makes his book easy to read. Complicated and interesting ideas are presented in a matter of fact way that is calming. But, as they say in B horror movies -- *too* calming. His style is deceptive, for it tends to draw the reader into a feeling of relief: Finally, an AI researcher has weighed in on the problem of consciousness. Now we can get beyond all that silliness about possible worlds, zombies, Cartesian intuitions, the neural correlate of consciousness, and dualism and all of its perverse variants. Now we can get down to it. McDermott is even so nice as to warn us that "a computationalist explanation of consciousness will inevitably sound like 'explaining away' rather than true explanation" (pp. 94-95). By this point in the book, the reader is inclined to say "Well of course, that just goes with the territory. Don't worry about that, just proceed with your theory." All the while, the not so obvious is made less obvious: perhaps there is no satisfying explanation of consciousness to be had. Perhaps, though there might be a *theory* , of sorts, about consciousness and its realizations, there will never be an *explanation*. Perhaps here, as for crucial parts of quantum mechanics, explanation and scientific theory pull apart. It is a dark and surprising fact of life in our universe that a useful scientific theory doesn't have to make the world more intelligible. Perhaps this is the best argument for realism that can be made.

McDermott's computational theory of conscious is of the "internal model" variety. All internal model theories of consciousness follow approximately the same route -- a route, by the way, that we have a lot of sympathy for. First, you argue that intelligent cognition at the human level requires an internal model of the system itself

which the system can access *as an internal model*. That is, the system knows the internal model is an internal model of it. (McDermott's cleverness as this stage is nicely exhibited in his discussion of our perception of a bent rod in water and our perception of having freewill.) Secondly, you establish another requirement: that via some kind of recursivity the system must itself be part of the internal model. That is, that the system needs a *self-model*, one built out of its internal model plus various symbols that refer to or denote itself (and not just parts of it). Now, at the third step, internal model theories have a choice. They can either wax poetical about the "emergence" of the self, hence self-consciousness, hence consciousness from the self-model, or they can take all the actually producible symbols, representations, and recursive access, which are always preceded with qualifiers "something like" or "nearly like" and sidle up to the real thing and then when no one is looking (often, not even the author), drop the qualifiers. The first branch always involves some mysticism. It is applauded by many because, after all, consciousness is the essence of being human and being human *is* mystical, or at least is should be because it is so marvelously wonderful; the step from machine to conscious entity had better involve something darn like magic. But McDermott will have none of this. He is valiantly nonmystical. He opts for the second branch. His story here is as good as one would want to read (see his chapter three for all the fascinating details). And one would really like him to succeed. But like someone trying to sail over the Grand Canyon on a motorcycle traveling at 100 miles per hour, and like all second branchers, McDermott doesn't have enough speed to make it to the other side -- to real consciousness, real qualia, real experience.

Of course, to this criticism, McDermott will simply say "Almost any materialist explanation, even the correct one, is going to have this problem [of not seeming true or correct] . . . because of the wide gulf between our intuitions about matter and our intuitions about mind" (p. 95). This clever ploy of making a virtue out of implausibility is disarming. Still, must do our best not to lose sight of the goal: satisfyingly explaining consciousness. But that is not in the cards. We believe the goal is unachievable. Hence, no theory, not even McDermott's, as he would be the first to admit, is going to explain consciousness and hence no theory will ever strike us as true. That is bad for a theory. There are other routes to truth, though. Perhaps McDermott's theory could be

made predictively adequate (this is the route followed by quantum mechanics). Then we could accept it without believing it. But as matters now stand, McDermott's computational theory of consciousness is neither believable nor predictively adequate.

The culprit for all of this is right in front of us. It is consciousness itself. It is not merely that there is a gulf between our intuitions about matter and our intuitions about mind. Science has a long history of building bridges over gulfs between seemingly incompatible intuitions. Our intuitions about continents don't allow them to move. Our intuitions about species don't allow them change. Our intuitions about motion and the sun don't allow it to remain still while we move around it. But nowadays, all these intuitions live happily together. Many hold out the same for consciousness. Sure there is a gap between our understanding, intuitions and all, of matter and mind, but just wait until next year -- when we all come to accept some internal model theory -- or the year after that. But consciousness, by its very nature, prevents us from satisfyingly explaining its material origins. Here, quite briefly, is an argument for this.

There is an intuition most of us have upon which dualists base their arguments, and which most materialists and all naturalists try hard to ignore. This is the intuition that our conscious experiences could be just what they are regardless of how the world is, that somehow our consciousness need not cohere with how the physical world actually is. We call this intuition our *Cartesian intuition*. Our Cartesian intuition is the kissing cousin of the *zombie intuition*: the intuition that all of our cognition could occur exactly like it does it us but in some other creature that was completely devoid of conscious experience. It is the zombie intuition that dualist frequently base their arguments on.

What really needs to be done is to explain *why* we have the Cartesian intuition in the first place -- a strategy McDermott also endorses. Suppose that our Cartesian intuition is due to consciousness itself. (The argument for this is complicated. See Dietrich and Hardcastle, 2005, and Dietrich and Gillies, 2001. For now, just adopt it as an assumption.) Having supposed this, it follows that explaining how consciousness arises due to material properties of neural processes is not possible. Here's why.

Consciously experiencing one's neural processes, say, by looking at them using some sort of sophisticated imagining technique, can only give one some bit of experience that one can now relate to the rest of one's experiences. It will never alleviate our Cartesian intuitions, which is what we need in order to have a satisfying explanation. Second branchers claim that to explain consciousness, all we need is to be able to draw the appropriate correlations between conscious experience and something physical. But that isn't enough. We can *agree* to draw a materialist or reductive inference using a form of inference to the best explanation, but the inference will never be compelling. Being the phenomenon that it is, consciousness *logically* prevents us from seeing how it could supervene on material. Hence, our intuitions tell us that it doesn't. (And for all that, perhaps it doesn't.) But even if dualism were true, that alone wouldn't cause us to have the Cartesian intuitions that we do. We have those intuitions not because of the truth of any "ism," but because we are conscious. And this renders suspect all theories of consciousness.

Since, presumably, all readers of McDermott's book are going to be conscious, his theory will strike none of them as true, either. And that is a shame. On the other hand, the correct view of the situation adumbrated above does leave an essential mystery at the heart of being human, and at the heart of being an intelligent machine. Perhaps, with McDermott, we could all accept that this doesn't devalue humans, but ennobles smart machines.

References

Dietrich, E. and Gillies, A. (2001). Consciousness and the limits of our imaginations. *Synthese* v. 126, n. 3, pp. 361-381.

Dietrich, E., and Hardcastle, V. G. (2005). *Sisyphus's Boulder: Consciousness, Science, and the Nature of Philosophy*. Amsterdam, The Netherlands: John Benjamins.