Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Artificial Intelligence 151 (2003) 227–235
Artificial Intelligence
www.elsevier.com/locate/artint

Book Review

Reading McDermott ☆

Varol Akman

Department of Computer Engineering, Bilkent University, Bilkent, Ankara 06800, Turkey

Executive summary

The author is interested in computational approaches to consciousness. His reason for working in the field of AI is to solve the mind-body problem, that is, to understand how the brain can have experiences. This is an intricate project because it involves elucidation of the relationship between our mentality and its physical foundation. How can a biological/chemical system (the human body) have experiences, beliefs, desires, intentions, and so on? Physicists have good reasons to persuade us that ours is a material world that obeys physical laws. Once we commit ourselves to this view, it sounds quite bewildering to think that there is a place for independently existing minds in such a world.

When physicists speak, McDermott listens. His hypothesis is that we are all contraptions designed by evolution. The concept of mind arises because people's brains are biological computers. Crudely put, minds are produced by brains. The essential contribution of this rich book is an extended argument about how one can be more specific about the way this production is realized (see Appendix A).

1. Introduction

The great American philosopher Donald Davidson once said [3, p. 14]:

☆ Review of: Drew McDermott, *Mind and Mechanism* (A Bradford Book, MIT Press, Cambridge, MA, 2001) xv+262 pages, US$32.95, ISBN 0-262-13392-X. A different version of this review has appeared in *Notre Dame Philosophical Reviews* (May 2002) as follows: http://ndpr.icaap.org/content/archives/2002/5/akman-mcdermott.html. Grateful acknowledgment is made to Gary Gutting, Editor-in-Chief of *NDPR*, for permission to use several portions of that version.

E-mail address: akman@cs.bilkent.edu.tr (V. Akman).
URL: http://www.cs.bilkent.edu.tr/~akman.

0004-3702/$ – see front matter
doi:10.1016/S0004-3702(03)00119-X

> If I were to discover that Daniel Dennett is made of silicon chips I would not change my mind about his mental powers, his feelings, or his intentions. Nor would I be greatly surprised.

McDermott thinks likewise. His major thesis is that people are machines, more specifically computers. People have minds because their brains (bio-computers) give rise to minds [p. 2]:[1]

> The biological variety of computer differs in many ways from the kinds of computers engineers build, but the differences are superficial. When evolution created animals that could benefit from performing complex computations, it thereby increased the likelihood that some way of performing them would be found. The way that emerged used the materials at hand, the cells of the brain. But the same computations could have been performed *using different materials, including silicon* (my emphasis).

In explaining how minds are produced by computers, McDermott cautions that one has to be careful. Positing that brains 'discharge' minds in some way, or suggesting that minds are emergent properties of brains (just like hardness is a property of diamonds) would not be entirely convincing. One needs to spell out in detail *how* computers can have minds. This is precisely what McDermott aims to show in *Mind and Mechanism* (*M&M* in the sequel).

2. The mark of the mental

John McCarthy likes to suggest, perhaps somewhat jokingly, that even devices as simple as thermostats have beliefs. According to him, a thermostat has, at any given time, one of the following beliefs: "It's hot in here", "It's chilly in here", and "It's cosy in here". But what is a belief anyway? How does it acquire the content it has (e.g., that it's *chilly*)? These questions cannot really be answered without clarifying the concept of 'a mechanism with a mind'. What conditions must be satisfied by a mechanism, say a robot, before we can attribute a mind to it?

In essence, this problem concerns the relation between mental and physical properties. After all, a robot is an inorganic electro-mechanical device, and it is possible to multiply the questions: Can it feel pain? Can it exhibit emotions like anger? Can it develop a taste for Paul Auster novels? Can it decide to join Greenpeace? Such questions have essentially been the subject matter of the philosophy of mind. A central problem of this discipline—first conceived by Descartes in its present form—is the *mind-body problem*.

McDermott sees mind as a self-fulfilling description, a description that brings mind into being. He offers, in Chapter 1, a fine analogy by considering the windows in the user interface of a personal computer. These windows are there because the computer is working in a way supporting their existence. It is able to do so by running procedures that

[1] Plain page references such as this one are to the work under review.

use data structures—those textual descriptions portraying what the windows are to be like. In a nutshell, a formal description starts to act *causatively* when interpreted by a computer. In some sense, the mind is also a window, albeit one with a Herculean description.

Chapter 2, the longest, is a masterly account of the state of the art in AI. Discussed in this chapter are such classical problems of AI as game playing (computer chess being the classic example), neural networks (including a clear discussion of the tensions between the symbolic vs. nonsymbolic approaches to AI), computer vision (optical flow), robotics (depth maps), speech recognition and natural language understanding (hidden Markov models), and automated theorem proving (mathematical reasoning). At first glance, this chapter looks like a stranger among the rest. After all, this is the only chapter that reports technological advances (along with their underlying theoretical foundations, to be fair). But the author has a good reason for including it: to draw a clear boundary between what we can now build and what is sheer speculation. Thus, when McDermott deliberates about computational mechanism in various other places in *M&M*, he in fact is referring to mechanisms like the ones mentioned in Chapter 2.[2] It is also in this chapter McDermott issues and defends intricate claims like [p. 46] "Computers don't deduce conclusions about things; they perform computations about them" or "If symbols do denote anything, it's because they are connected by the right kind of causal chains to the things they denote". The chapter is concluded with a fine account of creativity. McDermott thinks that creativity in one person need not resemble creativity in another person. He concludes his speculations with the following appraisal:

> [T]here may be a bit of randomness in the thought process of a creative person, but [this is] not the important part. The important part is a few key tricks for generating good ideas in the person's domain of expertise.

3. The ways things seem to us

McDermott attaches great importance to models of the world. These models include models of the self. The self-model is the source of everything one knows about oneself. More interestingly, it is because of the self-model that one believes in free will. Subtract the self-model and you would end up with an entity which is not a fully functioning person anymore.

For McDermott, *the problem of phenomenal consciousness*—how a physical object can have experiences—is the toughest nut to crack. To put the matter differently, explaining what it is to have qualia—the way things are experienced by conscious beings—is the hardest problem. To this end, McDermott first tries to demystify the notion of free will. His insight is that a robot must model itself in a different way from other objects in its world in order to avoid infinite regress. Consider a straightforward cycle of events in the life of

[2] Chapter 2 might very well serve as a fleeting introduction to AI, for Computer Science undergraduates. I hope that McDermott makes some version available in his home page so that instructors can assign it for reading. Before students delve into a bona fide AI text, this concise panorama of the AI landscape would greatly help them place things into proper context.

a robot, starting with perception, leading to making a tentative prediction, and concluding with revising an action. McDermott claims that this chain of events cannot be accurately embedded in a model unless the symbols denoting the robot itself are flagged as free from causality. Why is this so?

Let SHRDLU be a robot employing a model of the world in which it is situated. This model resides in the memory of the robot and includes many symbols, including a symbol ι, which SHRDLU uses to refer to itself. This is done in the least sophisticated way, carrying no philosophical connotations. For example, when SHRDLU moves its body from one place to another, it just enters this fact to the model, saying that now ι is at location such and such. Assume, for the sake of the argument, that something is fast approaching SHRDLU and has all the characteristics of a bull. The robot calculates that it may be hit by the bull and destroyed if it does not move away. (It is programmed to avoid damage to itself, along the lines of Asimov's laws.) By a circularity argument, it turns out that SHRDLU must model itself in a totally different way from other objects. The causal chain from perceiving the bull to tentative envisionment (of collision) to action revision (move away) cannot be fully represented in the model. This is due to the fact that the making of tentative predictions involves the model itself. Thus, the symbol ι must be declared as immune from causality, lest the modeling software of SHRDLU may fall into infinite regress.[3]

It may sound odd that in McDermott's proposal, the self plays such a simple role. However, the history of philosophy presents us with similar arguments about the reducibility of 'I' (the essential indexical) to more straightforward equivalents. Thus Strawson [9, p. 95]:

> [Wittgenstein] thought that there were two uses of 'I', and that in one of them 'I' was replaceable by 'this body'. So far the view might be Cartesian. But he also said that in the other use (the use exemplified by 'I have a toothache' as opposed to 'I have a bad tooth'), the 'I' *does not denote a possessor*, and that no Ego is involved in thinking or in having toothache; and referred with apparent approval to [the] dictum that, instead of saying 'I think', we (or Descartes) ought to say 'There is a thought' [...]

Geach [6, pp. 117–119] also makes a similar overture:

> Let us begin by reminding ourselves how "I" is used in ordinary life with psychological verbs. If P.T.G. says "I see a spider" or "I feel sick", people will ordinarily think that the speaker who says this, P.T.G., sees a spider or feels sick. [...] Now consider Descartes brooding over his *poêle* and saying: "I'm getting into an awful muddle—but who then is this 'I' who is getting into a muddle?" When "I'm getting into a muddle" is a soliloquy, "I" certainly does not serve to direct Descartes's attention to Descartes, or to show that it

[3] The situation is reminiscent of Tarski's well-known distinction between object language and metalanguage, and his regimented language hierarchy in the context of paradoxes or circular arguments [1, p. 5]. The metalanguage must be capable of expressing everything expressible by the object language plus have resources that go beyond the object language. An obvious consequence of this is a stratified system of languages. For a given language in the hierarchy, only talk about the truth of sentences in former languages, and not about the truth of its own claims, is possible.

> is Descartes, none other, who is getting into a muddle. We are not to argue, though, that since "I" does not refer to the man René Descartes it has some other, more intangible, thing to refer to. Rather, in this context the word "I" is idle, superfluous; it is used only because Descartes is habituated to the use of "I" [. . .] in expressing his thoughts and feelings to other people.

McDermott's approach to explaining free will can also be used to explain qualia. The key move is to notice that a sequence of goals must come to an end with a goal that can't be further questioned. Imagine SHRDLU on a dangerous mission to save people from drowning in a river. As long as the robot is immersed in water, its underlying system can label wetness as "undesirable but OK for the time being". According to McDermott, at this point the robot's apprehension of wetness is analogous to a quale of disagreeableness. In representing this state, SHRDLU classifies it as "to be shunned or evaded as much as possible". To cite another of McDermott's examples [p. 102]: "[A] robot may dislike going into burning buildings because it dislikes heat. But it doesn't dislike heat because of further bad consequences; high heat is *intrinsically* not-likeable" (my emphasis).

But then there is no need for qualia in the computational system of a robot. Consciousness arises through the employment of a self-model, and qualia are occasioned by the process of self-modeling. That means that what is manifested by our robot is virtual consciousness, which eventually boils down to physical events. While McDermott does not explicitly cite it, Dennett's concluding paragraph of his celebrated manifesto is even stronger in tone [4, p. 639]:

> So when we look one last time at our original characterization of qualia, as ineffable, intrinsic, private, directly apprehensible properties of experience, we find that there is nothing to fill the bill. In their place are relatively or practically ineffable public properties we can refer to indirectly via reference to our private property detectors—private only in the sense of idiosyncratic. [. . .] So contrary to what seems obvious at first blush, *there simply are no qualia at all* (my emphasis).

It is true that this computationalist explanation of consciousness has all the characteristics of 'explaining away' rather than a true explanation. But, this is expected. It is, after all, hard to see how a computational model, incorporating inputs, outputs, and—what else—computations, can have phenomenal consciousness. One's best bet in this case is to argue, as McDermott does, that something *like* consciousness will be exhibited by the system.

Chapter 4 is devoted to a thorough defense of the proposals (formulated in Chapters 1 and 3) against various (some already classical) objections. Discussed in Chapter 4 are numerous full-fledged or partial theories of consciousness or related matters by influential philosophers: (in no particular order) Michael Tye, Peter Carruthers, Andy Clark, Georges Rey, Ned Block, David Chalmers, Frank Jackson, Daniel Dennett, Sydney Shoemaker, Thomas Nagel, William Lycan, John Searle, et al. Almost all of McDermott's arguments and analyses in this chapter are to the point, upright, and illuminating. His

grasp of the contributions of the aforementioned authors is commendable in its clarity and objectivity.

Chapter 5 is somewhat technical and treats a question which is not central to McDermott's general theory: the observer-relativity of symbols and semantics. As a notable proponent of observer-relativity, John Searle famously said [8, pp. 208–209]:

> For any program and for any sufficiently complex object, there is some description of the object under which it is implementing the program. Thus for example the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements that is isomorphic with the formal structure of Wordstar. But if the wall is implementing Wordstar, then if it is a big enough wall it is implementing any program, including any program implemented in the brain.

Thus a symbol can mean anything and whether something is a computer depends entirely on whether a person regards (and uses) it as a computer. McDermott's refutation of this dangerously relativistic viewpoint is based on his painstaking analysis of the notion of *decoding* (a mapping from a computer's states to its computational realm). It is then argued that while everything might be considered as a computer sometimes (with respect to some decoding), this fact does not endanger the concept of a computer. Using a clever continuity argument (small state perturbations causing little difference in the output), McDermott shows that if someone makes the Searlean claim just mentioned, then the burden of proof is on him or her to demonstrate that the continuity requirement is not violated.

4. God and Man at Yale

Assuming that McDermott's general views regarding the self and qualia are correct, an infamous question raises its head: Is there anything precious or sacred about a person (= a machine)? The history of AI is replete with numerous instances of this question because people are understandably worried about the moral dimension of the AI enterprise. A related question that has been asked for the umpteenth time (most recently in a flashy Spielberg movie) is this: Can an artificial intelligence "be the subject or object of moral judgment?" [p. 217]. McDermott states that a robot can have phenomenal consciousness but lack morals. If I understand him correctly, his argument for this is based on the behavior of a program. Assume we build a robot which is just like an average person when it comes to attributes such as love, morality, humor, etc. Now, it is conceivable that by just tweaking some program segments, we could get radically different (perhaps unrecognizable) versions of these attributes. Consequently [p. 221],

> We'll take robot aesthetics [similarly, robot morals—VA] seriously if the robots ever reach a point where we can debate them, destroy them, or lobotomize them, but can't change their tastes [moral codes—VA] by adjusting a few lines of code.

The last big question that McDermott grapples with is belief in God in the absence of dualist non-physical realms.[4] His stance is not anti-religionist; he doesn't consider religion a menacing force. Neither does he regard belief in God as a potentially short-lived activity in the broad development of civilization. McDermott is convinced that mankind will always be troubled by finiteness and long for the comfort provided by the unfathomable. In a memorable passage, he writes [pp. 237–238]:

> The only way to reconcile God's silence with his existence is to assume that *he poured himself into the world when he created it*. His intervention in the world consists in creating laws of physics that allow complex physical systems to evolve that understand those very laws; in making it possible for weird creatures to evolve from monkeys and grasp the need for an overarching set of moral principles (my emphasis).

Thus, McDermott's God is one who does not answer prayers or does not prevent bad things from happening to good people. It is a God who turned "us loose in such an uncaring world" [p. 239]. According to McDermott, there is only one rational thing to do then: to take a reverent stance and accept the world as *given*.

5. Points of contact?

Do I have any suggestions for a future edition of *M&M*? Well, probably just these two. Firstly, except for a passing remark (on something else), McDermott does not connect his proposal with the doctrines of Davidson. This is a pity because I do believe that such a comparison would be promising, even fruitful. As some readers will know, Davidson has worked out an ingenious answer, known as *anomalous monism*, to the puzzle of the mental [2]. His is an informal but difficult argument and cannot be done justice in this brief outline. Basically, Davidson takes it for granted that the essential properties of matter as described by physicists are the only properties we have. Thus, he subscribes to some form of materialism. However, he thinks that one can be a materialist while also asserting that the mental cannot be reduced to the physical. Assume that you have complete knowledge in front of you of your brain and any relevant neurophysiological systems. According to Davidson, this knowledge cannot constitute knowledge of your beliefs, desires, intentions, etc. This he maintains without really taking a dualist stance, that is, without assuming that your mind has a separate kind of existence. Rather, his point is that our 'vocabulary' for describing the mental does not match the concepts of physics in the right way.[5] For example, he sees the principle of rationality as a most crucial aspect of the mental (especially belief), and holds that this principle has no echo in physical theory. Davidson's thesis is that the nature of mental phenomena does not permit law-

[4] *Historical aside:* William F. Buckley published *God and Man at Yale* precisely 50 years ago. In Buckley's opinion, Yale was originally founded upon the belief in Almighty. But at the time of his writing Buckley thought the university was cherishing atheism.

[5] On a similar note, John Searle once said that weather concepts such as 'partly cloudy' are not systematically related to the concepts of physics.

like regularities connecting the mental phenomena with physical events in the brain [3, p. 18]: "The mental and the physical share *ontologies*, but not, if I am right, classificatory *concepts*".

Secondly, I detect a likelihood for commerce between McDermott's account of the self and *pathologies of self*, as studied by Rom Harré in the context of discourse. In cases of multiple personality disorder (MPD) a commonsense principle that can be summed up as "one person per body and one body per person" seems to be seriously violated. Paraphrasing the characterization of American Psychiatric Association [7, pp. 155–156], MPD is the existence and taking full control of within the person two or more distinct personalities, each with its own pattern of perceiving and understanding the environment and self. Here is the famous case of Miss Beauchamp [7, p. 152]:

> Under hypnosis she began to address remarks to and about herself, as if from the point of view of someone else. Later, as her condition developed she would address comments from the point of view of yet another 'person'. Prince [her psychiatrist] called these 'speakers' BI, BII and BIII. BII began to take on personhood as a characteristic pattern of pronoun usage marked a complementarity of address between her and BI, Miss Beauchamp proper. The 'I'–'you' pair shifted indexical reference from Miss Beauchamp to her alter ego.

It turns out that pronouns played a key role in remedying Beauchamp's troubles. A cure was obtained [7, p. 153] "by the incorporation of the memories of each voice within a common autobiography, that is as a temporally coherent and continuous story as indexed by the pronoun 'I'. Tying some recollections to 'you' and 'she' had ceased".

6. Conclusion

In the preface of his collection of papers on language [5], Michael Dummett offers some useful advice. He hopes that readers might enjoy his work and be stimulated to new approaches of their own. More importantly, he says that he does not expect agreement. I think the same comments can be made for *M&M*. In the space of a mere 250 pages, McDermott spans the landscape of almost every important problem in computation, cognition and mind, while not causing indigestion (but presumably invoking lots of disagreement). A capable author in lucidly explaining the most intriguing phenomena, he formulates fine—but tentative—solutions for some of these vexed questions. One strength of this book is its almost encyclopedic coverage; it simply looks like the agenda of a busy philosopher of mind-language-and-morals. And while it may be surmised that in treating so many diverse concepts it would be easy to fall into the trap of superficiality, I did not really notice anything of the sort. McDermott has a natural gift for explaining the knotty or perplexing with such grace (and vivid examples) that one cannot help but admire his virtuosity. Thus, his worries about not "using the usual philosophical tools to approach [philosophical questions]" [p. 24] or not conducting his discussions "in the pure philosophical style" [p. 25] are unwarranted.

So, do you want an intuitive yet frequently deep chronicle of what is keeping philosophers of AI occupied nowadays? Do yourself a favor and read this book.

Appendix A. Chapters of *Mind and Mechanism*

- 1. The Problem of Phenomenal Consciousness
  The hard problem
- 2. Artificial Intelligence
  A survey of the state of the art
- 3. A Computational Theory of Consciousness
  A detailed explanation of McDermott's proposal
- 4. Objections and Replies
  Various objections to the aforementioned theory and McDermott's rebuttals
- 5. Symbols and Semantics
  Despite some claims, 'computer' and 'symbol' do not denote ill-defined notions
- 6. Consequences
  Implications of the theory in the ethical and religious realms

References

[1] J. Barwise, J. Etchemendy, The Liar: An Essay on Truth and Circularity, Oxford University Press, New York, 1987.
[2] D. Davidson, Mental events, in: D. Davidson (Ed.), Essays on Actions and Events, Clarendon Press, Oxford, 1980, pp. 207–225. (This paper was first published in 1970.)
[3] D. Davidson, Representation and interpretation, in: K.A. Mohyeldin-Said, W.H. Newton-Smith, R. Viale, K.V. Wilkes (Eds.), Modelling the Mind, Clarendon Press, Oxford, 1990, pp. 13–26.
[4] D.C. Dennett, Quining qualia, in: N. Block, O. Flanagan, G. Güzeldere (Eds.), The Nature of Consciousness: Philosophical Debates, MIT Press, Cambridge, MA, 1997, pp. 619–642. (This paper was first published in 1988.)
[5] M. Dummett, The Seas of Language, Clarendon Press, Oxford, 1993.
[6] P. Geach, Mental Acts, Thoemmes Press, Bristol, 1992. (A reprint of the 1971 edition.)
[7] R. Harré, The Singular Self: An Introduction to the Psychology of Personhood, Sage Publications, London, 1998.
[8] J.R. Searle, The Rediscovery of the Mind, MIT Press, Cambridge, MA, 1992.
[9] P.F. Strawson, Individuals: An Essay in Descriptive Metaphysics, Methuen, London, 1959.