Available online at www.sciencedirect.com
SCIENCE DIRECT®
Artificial Intelligence 169 (2005) 184–191
Artificial Intelligence
www.elsevier.com/locate/artint

Book review

Jeff Hawkins and Sandra Blakeslee, On Intelligence, Times Books, 2004.

Hawkins on intelligence: Fascination and frustration

Donald Perlis

University of Maryland, College Park, MD 20742, USA

Available online 21 October 2005

1. Introduction

I find *On Intelligence* to be both fascinating and frustrating. Fascinating in that it hints at a potentially unifying view (the memory-prediction model) of brain structure and function; and frustrating in that it is very sketchy and in that it also makes claims about AI and conscious experience that seem to me to be questionable. I will mostly leave the neural aspects to other reviewers, focusing my remarks primarily on four things: (i) general comments about the model and how it relates to AI, (ii) Deep Blue as case study, (iii) the Turing Test and its impact on AI, and (iv) consciousness.

But first some general comments on the book, to orient the reader. Jeff Hawkins (with co-author Sandra Blakeslee) has written an easy-to-read book on a topic of broad general interest—the nature of intelligence and how it is that brains have it and why machines don't yet—in which his own excitement comes through and infects the reader as well. The account is largely biographical, in which Hawkins (inventor of the Palm Pilot) describes how he came to his current ideas about artificial intelligence, the brain, and related topics including Deep Blue, the Turing Test and the Chinese Room. He also describes the ideas themselves in an engaging way, and ends the whole thing with eleven in-principle testable theses that his theory suggests. The upshot is that he thinks the

E-mail address: perlis@cs.umd.edu (D. Perlis).

doi:10.1016/j.artint.2005.10.012

memory-prediction model of the brain will revolutionize not only neuroscience but also the entire approach to building smart machines, and may make the latter possible within a few years.

It's a fun read, whether or not one thinks his ideas are new or correct. Yet it also may cause ire in some readers who may feel that Hawkins has misrepresented research trends in AI, especially in regard to the importance of the detailed nature of inner processing, and to the particular importance of perception, memory and prediction as key parts of that inner processing.

2. The memory-prediction model and AI

Hawkins' basic premise seems right: perceptual-memory-based predictions surely do play a fundamental role in intelligence and hence in brain structure and function. What does this mean? That an intelligent agent learns from experience, and in particular builds up a model of the world by perception (not only of what is out there but of what actions achieve what results); and that this experience is remembered and available virtually instantly for use in deciding what to expect next.

You might be thinking, "Ok, fine, who would disagree? This sounds like the mainstream AI view: perceive-learn-store-recall-infer-plan-act, the long-envisioned hybridization of various AI subspecialties into a single general-purpose intelligent agent, what we are all working toward". No, according to Hawkins: what is missing from mainstream AI is an appreciation for the special functional architecture of the brain; special in that it is designed (has evolved) with this hybridization, and especially for what he calls prediction: estimating what the agent would normally perceive next, given what it has perceived previously. Just what makes him think AI is so far from the mark on this is not made totally clear in the book, but it is true that we still seem to be far from the goal of anything remotely approximating the general-purpose flexibility of human-level intelligence, and he thinks that a more brain-informed approach that focuses on prediction will move things ahead quickly.

Yet while we are not *there* yet, I would argue that we are moving in that direction, and have been for some time. In particular, we are putting more and more perceptive/predictive memory/world-modeling aspects into machines. It is nice to have someone of Hawkins' prestige sounding the call for this kind of work, but simply stating it in a book does not really solve any problems. The devil is in the details, and it is not clear that Hawkins has anything to offer beyond what is already being done.

Thus many have studied predictive aspects of mind, in AI (e.g., the frame problem, the ramification problem, non-monotonic reasoning), in developmental and cognitive psychology (e.g., the appearance-reality distinction, metacognition), in neuroscience (e.g., efference copy). But his idea that prediction is *the* organizing feature that underpins everything neocortical (and hence everything cognitive?) is striking indeed. Nevertheless, the various bodies of work from several disciplines just mentioned

indicate that the larger cognitive science community is not blind to that aspect of mind.[1]

3. Deep Blue as case study

Hawkins cites Deep Blue as an example of AI gone wrong, in the sense of doing a narrow task extremely well (playing chess) but totally missing the boat with regard to flexible intelligence or understanding. Deep Blue has the right I/O, it produces the right chess moves in response to its opponent's moves, but it knows not what it does.

Hawkins says (page 20): "understanding cannot be measured by external behavior", but he then follows this (page 21) with "The only way we can judge whether a computer is intelligent is by its output, or behavior". The first of these quotes I doubt many (any?) AIers would disagree with.[2] However, the second statement seems outright bizarre. With computers—as opposed to humans—we *can* look inside to see the processing. In the case of humans we can only guess, since (so far at least) we are not able to see into another's thoughts, nor very far into the details of their brain processes. Indeed, Hawkins takes a look inside Deep Blue, in order to conclude that it does massive speedy blind look-aheads of millions of options and does not (like a human expert) decide on a certain small number of important ones to examine. We can all agree Deep Blue is not intelligent, does not have understanding, not because we cannot look inside but precisely because we *do* know what goes on inside it. Hawkins argues that, in principle, we might be able to program a computer to do what a human (or a human brain) does, but he says no one is doing this and no one can do this until we first know what the brain is doing.

Hawkins also cites the Chinese Room of Searle [8] as example of a program with no understanding (page 20). Of course, as Hawkins is well aware, the "room" is a thought experiment (no code exists, since it is an argument intended to apply to *all* programs). Yet Hawkins also says that he thinks that the problem with the room-'program' is that it really has no understanding because it fails to have the right inner processing: it is missing the right sort of remembering of its experiences and using them for predicting. But this is to miss the point of Searle's claim: that *no matter what* kind of information-processing a computer (program) does, it will *only* (at best) provide a *simulation* of a human (or the brain) and will necessarily leave out anything that can sensibly be called intelligence or

[1] I can hear Hawkins (perhaps along with some other readers) groaning, "Surely you can't be serious that the mentioned efforts in AI, and especially the research in nonmonotonic reasoning—that painstakingly formal work—is bringing us closer to real, flexible, intelligent agents, dynamically updating their view of the world"? Well, to some extent I sympathize with this, namely to the extent that I am trying to nudge that area (at least by example of my own work) in more realistic directions, such as including real-time aspects, memory aspects, and so on. But I think there is broad agreement that this is needed, and less consensus as to how. Whether features of the *neocortical neural substrate*, e.g., its six-layered architecture, has anything to offer AI, is unclear, although Hawkins obviously thinks it does, and I think few would insist that it does not.

[2] Except in a *prima facie sense*; see below on the Turing Test for more on this. Note, after all, that teachers routinely judge understanding by giving their students tests that measure I/O, not internal processing. But the assumption is that there is very special internal processing going on behind the scenes, in the students' brains.

understanding. Hawkins disagrees with this and thus—despite what he says—he does not really agree with Searle's argument.

But we can all agree on Deep Blue: it is brittle. It cannot play checkers, nor can it learn to do so. This sort of observation applies as well to most AI programs today, as has often been noted. And it is the basis for much ongoing work, to remedy the situation. It is, for instance, the basis for the *metacognitive loop* (MCL) proposal being explored by my own group at the University of Maryland [1,5]. Key to MCL is the idea of an expectation: what the system thinks likely to happen next, so that it can decide whether an anomaly has occurred, requiring special attention.

Expectations are of course based on the system's current world model, which itself is built out of experience in a very broad sense: perceptions but also inferences and other inputs to its KB. For instance, perception may reveal that some rough-surfaced objects on some rough-surfaced tables travel with the table when the latter is moved, and that some other smooth-surfaced objects on some smooth-surfaced tables tend to slide off the table. Now we are told that an object that we cannot see is smooth-surfaced and the table is also smooth; what do we predict? Inference is required here, not simply memory. We need to connect the symbolic information that the object is smooth with the remembered look of past objects and the remembered slippage of those.[3]

How might this help Deep Blue? Well, it should notice that the game is different from expectation, and initiate a learning process for the new game. Of course, Deep Blue cannot do this, but endowed with a suitable version of MCL, it should be able to do so (this is the sort of thing we are investigating).

So, Hawkins seems wrong about AI-blindness to internal (cognitive) workings (as in world-model-based predictiveness) vs I/O behaviorism. How about his other claim, that the specific neural architecture that the brain uses for this modeling is critical to intelligence? That remains to be seen, but I know of no one in AI who thinks we should deliberately ignore findings in neuroscience. The brain is the only example we have of an intelligent system, and we would be foolish indeed to ignore it.

Hawkins bases much of his complaint about AI (that what is missing is an appreciation for the internal processing—whether at the neural level, or simply at the cognitive level—that constitutes intelligence, beyond mere externally observable I/O behaviors) on Alan Turing, and his Turing Test, to which we turn next.

[3] One can make much the same point about, for instance, planning: in unfamiliar settings, planning is not a routine matter of recalling and applying a remembered sequence of past actions. Rather it requires fitting together a new sequence, out of parts that must be sorted and sifted from a large pool, and arranged in ways suited to the new setting. Yes, memory and prediction are very much involved, but in a way that amounts to heuristic search and inference; and nothing in Hawkins' model suggests new insights into how to do that. Even a trial and error approach (aka generate-and-test) requires mechanisms that lie beyond mere memory and prediction. Of course, to the extent that all information processing involves manipulating stored data (and hence, in a sense, memories) then all cognition is memory-processing; but this is a truism, so general as to be unhelpful in understanding any particular kind of processing.

4. The Turing Test and AI

Did Alan Turing—the father of computer science—get us off on the wrong foot in the famous Turing Test for intelligence?

Turing begins [12]:

> 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 and is expressed in relatively unambiguous words.

(Turing then goes on the describe the Test: can a machine successfully imitate a human in conversational give-and-take?)

Thus Turing explicitly states he is not offering a definition of thinking. He later refers to "the polite convention that everyone thinks", and suggests that we extend the same convention to machines that pass the Test. It seems to me that Turing is, in effect, taking the position that his Imitation Game is simply a *prima facie* test of intelligence, not a necessary and sufficient condition. That is, it may be a reasonable practical guide, if we find ourselves in need of making a decision on the matter, say for legal or other practical purposes, and in the absence of contravening evidence. What might such evidence be (that, despite passing the Test, a system is not intelligent)? One answer is obvious: the system turns out to have an enormous database of answers to everything, and simply performs a lookup whenever called upon to respond. This idea is one case of a program in which Searle [8] in the Chinese Room scenario can rightly argue his point, for instance (although it was around as folklore long before that). We would not regard such a system as intelligent: it does no thinking, has no understanding, no true flexibility to deal with the unanticipated (since everything has been anticipated for it in advance).[4]

But the Test does not tell us what intelligence really is—and Turing has distanced himself from that question. But then what good is his Test from a scientific perspective? As Turing says at the very end: "We can only see a short distance ahead, but we can see plenty there that needs to be done". His Test provides a challenge that can be worked on. And indeed some have taken up this challenge, e.g., contestants in the annual Loebner Prize Contest. But what has the impact of the Test on AI really been? Yes, it is famous, but many in AI have questioned it[5] and it appears that few in the relevant subfield of natural-language

[4] Presumably such a database would have to be large beyond any conceivable practicality, or even infinite, to work as stated, and this supports the Test as a practical guide (i.e., how else could the system be designed to actually work in real time and real space, except as one that does thinking-style processing), but that is another issue.

[5] See, for example, [4], for a particularly strong set of criticisms.

processing (NLP) are influenced by it in their day-to-day work. On the contrary, much of such work is highly informed by studies in linguistics and developmental psychology and even occasionally neuroscience.

5. Conscious experience

Hawkins suggests that ordinary consciousness is "what it feels like to have a neocortex" (page 194) and also amounts to "forming declarative memories" (page 196). He does not attempt to say how these two notions are related.[6] He also mentions qualia, and states, incorrectly, that this is a distinct kind of consciousness or a distinct issue or problem. I will comment on all three of these.

The problem of qualia, or qualitative states, is not a separate problem, but is simply a way of dramatizing the basic problem of (phenomenal, experiential) consciousness, aka subjective awareness.[7] That is, such states are characterized by having qualities, a something-it-is-like to be in that state.[8] It is like something to be awake; it is like something to dream; it is not like something to be in dreamless sleep, or to be dead, or to be a rock. These things-it-is-like-to-be, these subjective qualities, these sentient feels of being something, are the qualia. Some of them are color-qualia, or pain-qualia, but that is a detail, not the real point of the so-called problem of qualia. Hawkins hits on it when he raises the issue (page 198)—but then says no more about it—as to why there is any sort of qualia sensation, any sort of subjective feeling, in the first place. *That* is the problem![9]

Block [2] has described what he calls access consciousness, distinct from phenomenal consciousness; the former seems close to Hawkins' notion of declarative memories, and does not require any form of experienced subjectivity: computers have access consciousness of their memories. But phenomenal consciousness is precisely what philosophers tend to call *the* problem of consciousness, what Chalmers [3] calls the hard problem: what is it to be aware, to feel, to experience as a subject, to be something it-is-like-to-be. This hides two questions: (i) what is it, in physical (or brain or neural or other terms) and (ii) why is it so: what is it about that physical/brain/neural whatever-it-is, that makes it be aware, a self, a something-to-itself, a subject, an experiencer? Of course, only when we answer the second question will we be sure of the answer to the first. Nothing I can find in *On Intelligence* even hints at an answer to either question.

Hawkins describes declarative memories as "memories you can recall and talk about". But surely any computer can form declarative memories in the sense of stores of data,

[6] Presumably he would say a neocortex helps us form declarative memories, but that says nothing about why there is any feel to it; nor does it address whether, say, chimpanzees—who presumably do not form declarative memories, but who do have well-developed neocortices—are conscious or feel like anything (presumably they do).

[7] This is a fairly common understanding of the term "qualia", although some prefer a more narrow usage. See Searle [9] for a concurring opinion here.

[8] This very helpful way of characterizing subjective consciousness is due to Thomas Nagel [6].

[9] The literature on this is vast; my own attempt at an answer is given in [7]; roughly I urge the idea of a primitive ur-quale on which all other qualia are based, and I situate the ur-quale in a rather strong condition of self-representation.

even records of what has happened (as any operating system does) and later retrieve those records and offer them up to a human. Not in English, to be sure, but why should that matter? For that matter, some computer systems can even offer up reports in English, albeit a limited form of English. The SNePS system, or its various applications known as CASSIE [10,11], at the University at Buffalo, for instance, can report, in English, on what it has been doing, when asked. But I have heard no one argue that CASSIE is conscious, nor do I think Hawkins would so argue. It is, rather, that when *we* report on our doings, we are doing more than retrievings and reportings. We are also experiencing our acts, we feel like something as we engage in the conversation; and we also continue to feel like something when the conversation has been long over—the retrieving and reporting is not what gives us the sense of consciousness. Rather it is a matter of being aware, not in dreamless sleep; it is like something to be so. Hawkins alludes to this but prefers his declarative memory version without saying why, or how it differs from "awareness". And why the *forming* of the memory should be associated with consciousness is also not explained.

Hawkins gives a thought experiment here (pages 196–197). Suppose your memory of playing a tennis game is erased afterward; he rightly claims that you would say it never happened. Now if a videotape convinces you it did happen, you might then say you must have been playing while unconscious (as Hawkins argues), or you might just as well say that you have lost your memory of that time-period (a more likely result, as we all have had memory-lapses, but few have had episodes of unconscious game-playing). In any case, even accepting Hawkins' version, his apparent conclusion that this shows the presence of the memory is equivalent to having been conscious at the time, does not follow.[10] Thus by the very assumptions of the story, it would seem you were indeed conscious at the time of the game, and just because later on you have a false belief about it does not change the earlier fact of consciousness. Yes, of course, what you later *say* is that you had not been conscious during the game, but you are mistaken. So all this shows is that, as with most beliefs about what has transpired earlier, we can be mistaken about whether we had been conscious at an earlier time. The memory was erased, but not the consciousness that had occurred during the game—it is not that, later on, it is no longer true that you had been conscious while playing.

So, the thought experiment seems to show the opposite of what Hawkins claims: the presence of declarative memory is not the same as the fact of the consciousness: one can be removed, while the other cannot. To be sure, one is not conscious now of having played the game earlier, one is not conscious now of those (erased) memories. So, retention and recall of those memories at a later time are what constitute consciousness of them at that

[10] It is a little hard to be sure just what his claim is; he says (of your supposed assertion not to have been conscious during the game) "Therefore this meaning of consciousness is not absolute. It can be changed after the fact by memory erasure". That claim seems false—the fact of the game-time consciousness seems given in the description of the events—but he also says a few lines earlier "Your belief that you were conscious disappeared only when your declarative memory was erased". This latter seems correct enough. He might be saying that there is no fact of the matter as to whether one is conscious, it is just a matter of what one is willing to say about it, and that depends on what one remembers. But that raises two problems: (i) it suggests that there is nothing to be explained here, which flies in the face of his own claims elsewhere that consciousness has a feel to it; and (ii) it depends on your insisting you were not conscious, rather than saying you are not sure, or that you must have forgotten.

later time. But note that one has not become unconscious in the present, just unconscious of those events. Hawkins appears to conflate being in a conscious state with being conscious of a particular memory. The record of having been conscious is gone, just as one's record of making a bank deposit might be lost, but that does not mean one was not conscious or that no deposit was made, or that one is unconscious now.

There may be a one-way implication: consciousness may well require some sort of memory formation—after all, consciousness surely involves some sort of processing of (stored) information, hence of memories in some sense; but not the other way around: memory-formation (even declarative ones) can occur without consciousness.[11] Can a human utter apparently quite meaningful sentences—even ones that are true and that relate past events—while utterly unconscious (e.g., while in dreamless sleep, or in a coma)? Probably so; certainly computers can (not in dreamless sleep or a coma, but nevertheless unconscious, since they are never conscious at all, so far).

6. Conclusions

A good many workers in AI will probably find themselves sharing—as I do—most of Hawkins' positive claims, e.g., the need to pay attention to human cognition and brain function, the importance of a predictive world model learned from perceptual experience, the need for general unifying theories. But his negative claim—that these have been ignored by scientists—appears not to be as much on the mark, and he runs roughshod over a bunch of complex issues. Nevertheless, his highly readable and engaging book, and his prominence as a top-rate computer engineer/inventor, may help foster more public support (especially in the form of more funding) for this important research direction.

References

[1] M.L. Anderson, D.R. Perlis, Logic, self-awareness and self-improvement: The metacognitive loop and the problem of brittleness, J. Logic Comput. 15 (1) (2005).
[2] N. Block, On a confusion about a function of consciousness, Behavioral and Brain Sciences 18 (2) (1995) 227–287.
[3] D.J. Chalmers, Facing up to the problem of consciousness, J. Consciousness Stud. 2 (3) (1995) 200–219.
[4] P. Hayes, K. Ford, Turing test considered harmful, in: Proc. 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995.
[5] K. Hennacy, N. Swamy, D. Perlis, RGL study in a hybrid real-time system, in: Proc. IASTED NCI 2003, Cancun, Mexico.
[6] T. Nagel, What is it like to be a bat?, Philos. Rev. 83 (1974) 435–450.
[7] D. Perlis, Consciousness as self-function, J. Consciousness Stud. (1997).
[8] J. Searle, Minds, brains, and programs, Behavioral and Brain Sciences 3 (1980) 417–424.
[9] J. Searle, The Mystery of Consciousness, The New York Review of Books, New York, 1997.
[10] S. Shapiro, The CASSIE projects: An approach to natural language competence, in: J. Siekmann (Ed.), Lecture Notes in Artificial Intelligence, vol. 390, Springer-Verlag, Berlin, 1989, pp. 362–380.
[11] S.C. Shapiro, Embodied Cassie, in: Proc. AAAI 1998 Fall Symposium on Cognitive Robotics, 1998.
[12] A. Turing, Computing machinery and intelligence, Mind 59 (1950) 433–460.

[11] Unless of course these are *conscious* memories we are talking about—which is probably what Hawkins has in mind—but that begs the entire question!