EXPERT BOOK REVIEWS

Technical editor: R. Bhaskar, IBM T.J. Watson Research Center
Vendor contact: Steve H. Wilcox, *IEEE Expert*, 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, CA 90720-1264

The physical and mental bases of thought, and the impending death of closet dualism

How to Build a Person: A Prolegomenon

Reviewed by Mahzarin R. Banaji, Yale University

It has become especially fashionable of late to give voice to one's opinion about the infeasibility of strong AI. The sometimes mysterious objections converge to declare one impossibility: of building human-like intelligence. One critic is philosopher John Searle,[1] whose frequently repeated Chinese-room anecdote (claimed to disprove the merit of the Turing test) arms with equal ease professor and undergraduate alike when they must roll up their sleeves and defend the inimitability of pure human intelligence.

Among other objectors to artificial-intelligence building are philosopher Hubert Dreyfus,[2] mathematician James Lighthill,[3] physicist Roger Penrose,[4] and computer scientist Joseph Weizenbaum.[5] I mention their disciplines because it strikes me (a social psychologist interested in mental functioning) as a curiosity worth examining that most of the well-known criticisms of strong AI have been proposed by humanists and scientists whose primary expertise is not AI. On the one hand, it is obviously exciting to see such cross-disciplinary fire, especially at a time of increased disciplinary insularity. The criticisms of strong AI are many, some are probably legitimate, and addressing them can only benefit the field. On the other hand, it is somewhat alarming that many of the criticisms either imply or advocate the termination of the AI enterprise. Some critics have argued for and been successful at cutting off funding for AI research.[3] I refer to the following two criticisms because of their pervasiveness:

(1) *AI has failed to meet its goal.* Must an infant discipline be required to set, much less meet, a goal (the goal, of course, being the triviality of building intelligence from scratch)? This criticism reminds me of other austere requirements that societies and organizations impose on their citizens, such as standardized tests of cognitive abilities in grade school, or prescribed publication rates in universities (which exacerbate information pollution). Citing a failure to produce results within a specific time is always troublesome; a sobering exercise for such a critic might be to undertake a short essay about the contributions of his or her own discipline in its first 41 years.

The criticisms of strong AI are many, some are probably legitimate, and addressing them can only benefit the field. On the other hand, it is somewhat alarming that many of the criticisms either imply or advocate the termination of the AI enterprise.

Of course, a science is at all times accountable to the society it serves. But my amazement derives from the ease and certainty with which some critics divine the moment of judgment (although these critics are probably responding to expectations generated by some AI researchers). It is particularly ironic when cushion-chaired philosophers define when another discipline ought to pay up, since it is philosophy that has "lost" entire subareas because sciences have come along and taken over inquiry.

Because their intellectual makeup resembles those of scientists and scholars in other disciplines, I must assume that AI practitioners will accomplish comparable goals amidst the usual hill climbing that characterizes all scientific research. My position, apparently divergent from some critics, is this: A new discipline ought to be allowed to continue at the pace its own community finds appropriate, along paths regarded by its own community to be worthwhile. Bizarre requirements of intellectual output over a specific time are at best a hindrance to the freedom of scientific pursuit, and older, imperialist disciplines should at least be embarrassed to make such demands on newer ones.

(2) *Strong AI is not possible.* Not all critics of AI set up kitchen timers. Instead, they labor to document how strong AI (at least as it is conceived today) is wrongheaded or just plain mythical. From the stories about the lives of scientists my mother read to me, I quickly abstracted one fact: Important discoveries often occur in environments hostile to the ideas and amidst warnings of the impossibility of the enterprises. Occasionally, a criticism is mighty enough to cause a serious

setback: Kelvin successfully rejected Hutton's thesis of the age of the earth, which contradicted the date favored by Kelvin himself.[6] Such history should make us wary of the vigorous rejection of any new idea.

Large-scale technological projects are continuously scrutinized and questioned (such as the human genome and supercollider projects), and AI should be no exception. But in the absence of any serious impossibility proofs, what is it about the possibility of creating artificial intelligence that so unnerves and provokes? As compared to in-house debates about the feasibility of scientific projects, denials of strong AI often come from those who are safely away from the trenches of AI. This is not the occasion to examine the source of the criticisms in detail, so I will only mention two possibilities, the second of which leads to the topic of the book under review.

Discomfort with strong AI might reflect a special variety of ego bias. We have anecdotal, observational, correlational, and experimental evidence for various levels of ego bias (such as group-serving biases: racism, sexism, "religionism," ageism, nationalism, communalism, casteism, and species-centrism). Common to these is a certain social solipsism, a difficulty in accepting the existence of socially discrepant organisms, marked by the tendency to place undue importance on any kind of difference and to convert that difference into inequity. Criticisms of AI can be viewed as a form of species-centrism, of resistance to imagining or accepting a new species that might be human-like in both a superficial and a deep sense.

Because the debate today concerns the mere possibility of AI, it is concerned largely with issues of scientific and technological feasibility. If, however, artificial intelligence that equals (or duplicates) human intelligence is created, the issues will become legal, political, economic, social, and moral. Two thousand years from now, when a historian begins to write a thesis about the first discussions of strong AI, it may well be struck by the similarity between resistance to granting personhood to artificially intelligent organisms and the constitutional denial of personhood on grounds of race in the US in 1787.[7]

Issues of personhood bring us directly to *How to Build a Person: A Prolegomenon*, in which John Pollock admirably sweeps aside the critics of strong AI by not even addressing them, and instead gets on with the job of offering a serious and intellectually uplifting proposal about how to "build a person." Pollock's proposal, stated boldly and succinctly in the preface, is "a defense of three theses: token physicalism, agent materialism, and strong AI." His position on dualism comes through persuasively, concluding in a chapter that is as enjoyable as its delightful title — "Cognitive Carpentry" — suggests. Pollock's major contribution is in demystifying the relationship between the physical and mental. By discussing this issue in the practical context of building intelligence, Pollock undercuts the befuddled dualism that characterizes thinking about the mind-body problem.

Whether Pollock's specific experiment with building a person is successful or not, it is his strategy of "building" that not only satisfies our technological-empiricist hearts, but also will serve as prima facie evidence in support or refutation of challenges to strong AI.

(I had believed until recently that any form of dualism must be abhorrent to the 20th-century scientist, so I was surprised to encounter closet — perhaps even unintended — dualism in arguments about the uniqueness of consciousness, the even greater uniqueness of the unconscious, the mysticism of self knowledge, the baffling nature of emotion, and the enigma of cognition. Closet dualism emerges because of our relative lack of knowledge about and expertise with mental phenomena, which is why scholars from other disciplines, and even those in this field, are sometimes confused. Perhaps the newer sciences of the mind have not yet convinced older disciplines of the specialized nature of their questions, the expertise of their knowledge, and the importance of their methods. Why else is it that in our spare time we don't write books about quantum theory?

(The diffidence with which physicists and philosophers approach the mental world is understandable. It is a vast and terrifying thing, although no more so than the nature of the universe or the state of the world's economy. But it has ceased to be a purely philosophical problem for some of us since procedures have become available that allow empirical confirmation and replication. Parochial as this sounds, I believe experimental and other empirical evidence will ultimately remove the remnants of dualism that still hover around discussions about human thought.)

Pollock's book begins with a fable about a species, the Oscarites,* which we quickly realize is our own. Through them, Pollock introduces the nuts-and-bolts of his attempt to implement a person: external sensors for perception and pain (sense organs); internal, introspective sensors that sense the operation of the external sensors; and second-order perceptual sensors that make Oscarites aware or conscious of lower-order sensors. It is this awareness that allows an intelligent machine to "invent a mind/body problem." Pollock believes that awareness at varying levels of mental functioning is critical — indeed, it is the necessary and sufficient condition — for producing artificial intelligence. As his treatment of token physicalism and agent materialism unfolds, mental events are seen as physical events, and people (cognitive agents) are seen as physical

* Accuracy in scientific observation, inference, and writing fails when gendered language is used. The only annoying aspect of Pollock's book is his pervasive use of "masculinist" language ("he" and "man" most noticeably). The name "Oscar" is itself a striking example of just how fundamental is the social category of gender. Pollock uses the name to refer to all current humans, as well as to a species that will presumably have no use for the primary function of gender: reproduction.

objects. Neither claim denies the existence or importance of mental states and cognitive agents, but this approach does clear away the cobwebs that surround discussions of the relationship between the mental and the physical of human thought:

> It is important to stress the ordinariness of the kind of physical objects that are people. They are, of course, very complicated physical objects in that they have extraordinarily complicated structures, but there is nothing metaphysically extraordinary about them.

The physical and the mental are quite distinct — "How *could* the sensation be the same as the feel of the sensation? That would be analogous to a rock being the same as the feel of the rock." — and yet they are involved in a "single causal nexus" because they are "just physical events that can be sensed in a second way."

I find this reasoning enormously satisfying, partly because it provides an insight into just why the confusion of dualism mistakenly emerges. Using Pollock's analogy, imagine holding a ball (creating a tactile experience of sphericity) versus seeing the ball (creating a visual experience of sphericity). It is obvious that the same situation is experienced in two modes, and no confusion about the identity of visual and tactile objects emerges in our belief (independent of whether sight and touch actually have a common object). Likewise, Pollock argues that the epistemological justification for token physicalism is "almost precisely the same," and that the functioning of the hypothetical Oscarites "should convince us that it is at least possible that our internal senses are sensing events that can in principle also be sensed by our external senses and would, in later guise, be called `neurological events.' There is nothing logically absurd about this." So why the confusion? Again, Pollock's own words provide the best explanation:

> The visual/tactile isomorphism is very obvious to us because we are constantly aware of both sides of the isomorphism and we make continual use of it in our judgments about the world. On the other hand, the physical side of the presumed mental/physical isomorphism is buried deep within our bodies and is hard to get at.... These two pieces of reasoning must stand or fall together.... If we had transparent heads and the neurological events corresponding to the mental events were macroscopically visible, I think we would regard token physicalism as being as much of a commonplace as is the observation that what we sense proprioceptively is our own bodily movements.

If this is indeed a possible reason for the dualism that blurs the vision of otherwise astute professionals, I would guess that psychophysicists and psychobiologists (scientists who must probe the relationship between the physical and the mental with a unique evenhandedness) are least likely to fall prey to the dualism mystique.

In the next several chapters, Pollock elaborates on several constructs, relationships among them, and their link to actual implementation: the notion of a self as a necessary condition for rational thought; the importance of functional analyses of human cognition; the syntactical and content properties of thought; and finally, the conclusion that makes Pollock's version of a person unique: "The concept of a person must simply be the concept of a thing having states that can be mapped onto our own in such a way that if we suppose the corresponding states to be the same, then the thing is for the most part rational."

A thorough evaluation of the final chapter on "cognitive carpentry" lies outside the scope of this review. Whether Pollock's specific experiment with building a person is successful or not, it is his strategy of "building" that not only satisfies our technological-empiricist hearts, but also will serve as prima facie evidence in support or refutation of challenges to strong AI. Given the preliminary nature of Pollock's implementation of cognitive carpentry (a term I wish I had coined), it is his unclouded and pragmatic expression of the mind-body problem that will make a lasting contribution to this debate.

How to Build a Person: A Prolegomenon by John Pollock, MIT Press, Cambridge, Mass., 1989, ISBN: 0-262-16113-3.

References

1. J. Searle, "Minds, Brains and Programs," in *Mind Design*, John Haugeland, ed., MIT Press, Cambridge, Mass., 1981.
2. H.L. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence*, Harper Colophon, New York, 1972.
3. J. Lighthill, "Artificial Intelligence: A General Survey," in *Artificial Intelligence: A Paper Symposium*, Great Britian Science Research Council, London, 1973.
4. R. Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford University Press, Oxford, England, 1989.
5. J. Weizenbaum, *Computer Power and Human Reasoning*, W.H. Freeman, San Francisco, Calif., 1976.
6. J.D. Burchfield, *Lord Kelvin and the Age of the Earth*, Chicago University Press, Chicago, 1975.
7. E.C. Smith, *The Constitution of the United States*, Harper & Row, New York, 1979.

Mahzarin R. Banaji is an assistant professor in the Department of Psychology at Yale University. She can be reached at PO Box 11A, Yale Station, New Haven CT 06520-7447; e-mail, mbanaji@yalevm.ycc.yale.edu