Machine Intentions

Don Berkich
Department of Humanities
Texas A&M University - Corpus Christi
6300 Ocean Drive
Corpus Christi, TX 78412
berkich@gmail.com

15 July 2018

Abstract

Debate between philosophical skeptics and researchers in artificial intelligence has been rich, penetrating, and fruitful, if frustrating at times. Rarely speaking past one another, the skeptics set conceptual markers drawn from the philosophy of mind to caution against reading too much into the successes of AI. Yet claims on both sides are prone to excess. For instance, philosophical skeptics of developments in robotics have found their rarely contested assumption that a robot can act of its own accord puzzling: Why should we think that a mere *artifact*, no matter how complicated, could ever have the capacity to act of its own accord given that its purpose and function is completely determined by its design specification? The skeptic's intuition is that machine agency is deeply incompatible with machine-hood in just the way it is not with person-hood. Thus the actions of situated robots like the Mars rovers cannot be more than a mere extension of the roboticist's agency inasmuch as the robot's design tethers it to the roboticist's intentions. In this essay I analyze an early, albeit important skeptical marker to plea for moderation on both sides.

Introduction

There is a conceptual tug-of-war between the AI crowd and the mind crowd.[1] The AI crowd tends to dismiss the skeptical markers placed by the mind crowd as unreasonable in light of the range of highly sophisticated behaviors currently demonstrated by the most advanced robotic systems. The mind crowd's objections, it may be thought, result from an unfortunate lack of technical sophistication which leads to a failure to grasp the full import of the AI crowd's achievements. The mind crowd's response is to point out that sophisticated behavior alone ought never be taken as a sufficient condition on full-bore, human-level mentality.[2]

I think it a mistake for the AI crowd to dismiss the mind crowd's worries without very good reasons. By keeping the AI crowd's feet to the fire, the mind crowd is providing a welcome skeptical service. That said, in some cases there are very good reasons for the AI crowd to push back against the mind crowd; here I provide a specific and, I submit, important case-in-point so as to illuminate some of the pitfalls in the tug-of-war.

It can be argued that there exists a counterpart to the distinction between *original intentionality* and *derived intentionality* in agency: Given its design specification, a machine's agency is at most derived from its designer's original agency, even if the machine's resulting behavior sometimes surprises the designer. The argument for drawing this distinction hinges on the notion that intentions are necessarily *conferred* on machines by their designers' ambitions, and intentions have features which immunize them from computational modeling.

In general, skeptical arguments against original machine agency may usefully be stated in the *Modus Tollens* form:

	1.	If X is an original agent, then X must have property P.	
	2.	No machine can have property P.	
∴	3.	No machine can be an original agent.	1&2

The force of each skeptical argument depends, of course, on the property P: The more clearly a given P is such as to be required by original agency but excluded by mechanism the better the skeptic's case. By locating property P in intention formation in an early but forcefully argued paper, Lynne Rudder Baker[3] identifies a particularly potent skeptical argument against original machine agency. I proceed as follows. In the first section I set out and refine Baker's challenge. In the second section I describe a measured response. In the third and final section I use the measured response to draw attention to some of the excesses on both sides.[4]

1. With apologies to BBC Channel 4's "The IT Crowd", airing 2006-2010)

2. Consider John Searle's article in the February 23, 2011 issue of the Wall Street Journal, aptly entitled, "Watson Doesn't Know It Won on Jeopardy!"

3. L.R. Baker, "Why Computer's Can't Act," *American Philosophical Quarterly* 18 (1981): 157–163.

4. This essay is intended in part to serve as a respectful homage to Lynne Rudder Baker, whose patience with unrefined, earnest graduate students and unabashed enthusiasm for rigorous philosophical inquiry wherever it may lead made her such a valued mentor.

The Mind Crowd's Challenge: Baker's Skeptical Argument

Roughly put, Baker argues that machines cannot act since actions require intentions, intentions require a first-person perspective, and no amount of third-person information can bridge the gap to a first-person perspective. Baker[5] usefully sets her own argument out:

A	1.	In order to be an agent, an entity must be able to formulate intentions.	
	2.	In order to formulate intentions, an entity must have an irreducible first-person perspective.	
	3.	Machines lack an irreducible first-person perspective.	
∴	4.	Machines are not agents.	1,2&3

Baker has not, however, stated her argument quite correctly. It is not just that machines are not (original) agents or do not happen presently to be agents, since that allows that at some point in the future machines may be agents or at least that machines can in principle be agents. Baker's conclusion is actually much stronger. As she outlines her own project, "[w]ithout denying that artificial models of intelligence may be useful for suggesting hypotheses to psychologists and neurophysiologists, I shall argue that there is a radical limitation to applying such models to human intelligence. And this limitation is exactly the reason why computers can't act."[6]

Note that 'computers can't act' is substantially stronger than 'machines are not agents'. Baker wants to argue that it is impossible for machines to act, which is presumably more difficult than arguing that we don't at this time happen to have the technical sophistication to create machine agents. Revising Baker's extracted argument to bring it in line with her proposed conclusion, however, requires some corresponding strengthening of premise A.3, as follows:

B	1.	In order to be an original agent, an entity must be able to formulate intentions.	
	2.	In order to formulate intentions, an entity must have an irreducible first-person perspective.	
	3.	Machines necessarily lack an irreducible first-person perspective.	
∴	4.	Machines cannot be original agents.	1,2&3

Argument B succeeds in capturing Baker's argument provided that her justification for B.3 has sufficient scope to conclude that machines cannot in principle have an irreducible first-person perspective. What support does she give for B.1, B.2, and B.3?

B.1 is true, Baker asserts, because original agency implies intentionality. She takes this to be virtually self-evident; the hallmark of original agency is the ability to form

5. Baker, "Why Computer's Can't Act," p. 157.
6. Ibid.

intentions, where intentions are to be understood on Castaneda's[7] model of being a "dispositional mental state of endorsingly thinking such thoughts as 'I shall do A'."[8] B.2 and B.3, on the other hand, require an account of the first-person perspective such that

- The first person perspective is necessary for the ability to form intentions; and
- Machines necessarily lack it.

As Baker construes it, the first person perspective (FPP) has at least two essential properties. First, the FPP is irreducible, where the irreducibility in this case is due to a linguistic property of the words used to refer to persons. In particular, first person pronouns cannot be replaced with descriptions *salve veritate*. "First-person indicators are not simply substitutes for names or descriptions of ourselves."[9] Thus Oedipus can, without absurdity, demand that the killer of Laius be found. "In short, thinking about oneself in the first-person way does not appear reducible to thinking about oneself in any other way."[10]

Second, the FPP is necessary for the ability to "conceive of one's thoughts as one's own."[11] Baker calls this 'second-order consciousness'. Thus, "if X cannot make first-person reference, then X may be conscious of the contents of his own thoughts, but not conscious that they are his own."[12] In such a case, X fails to have second-order consciousness. It follows that "an entity which can think of propositions at all enjoys self-consciousness if and only if he can make irreducible first-person reference."[13] Since the ability to form intentions is understood on Castaneda's model as the ability to endorsingly think propositions such as "I shall do A", and since such propositions essentially involve first-person reference, it is clear why the first person perspective is necessary for the ability to form intentions. So we have some reason to think that B.2 is true. But, apropos B.3, why should we think that machines necessarily lack the first-person perspective?

Baker's justification for B.3 is captured by her claim that "[c]omputers cannot make the same kind of reference to themselves that self-conscious beings make, and this difference points to a fundamental difference between humans and computers–namely, that humans, but not computers, have an irreducible first-person perspective."[14] To make the case that computers are necessarily handicapped in that they cannot refer to themselves in the same way that self-conscious entities do, she invites us to consider what would have to be the case for a first person perspective to be programmable:

a) FPP can be the result of information processing.

7. H-N. Castaneda, *Thinking and Doing: The Philosophical Foundations of Institutions* (Dordrecht: D. Reidel Publishing Co., 1975).
8. Baker, "Why Computer's Can't Act," p. 157.
9. Ibid.
10. Ibid., p. 158.
11. Ibid.
12. Ibid.
13. Ibid.
14. Ibid., p. 159.

b) First-person episodes can be the result of transformations on discrete input via specifiable rules.[15]

Machines necessarily lack an irreducible first-person perspective since both (a) and (b) are false. (b) is straightforwardly false, since "the world we dwell in cannot be represented as some number of independent facts ordered by formalizable rules."[16] Worse, (a) is false since it presupposes that the FPP can be generated by a rule governed process, yet the FPP "is not the result of any rule-governed process."[17] That is to say, "no amount of third-person information about oneself ever compels a shift to first person knowledge."[18] Although Baker does not explain what she means by "third-person information" and "first person knowledge," the point, presumably, is that there is an unbridgeable gap between the third-person statements and the first-person statements presupposed by the FPP. Yet since the possibility of an FPP being the result of information processing depends on bridging this gap, it follows that the FPP cannot be the result of information processing. Hence it is impossible for machines, having only the resource of information processing as they do, to have an irreducible first-person perspective.

Baker's skeptical challenge to the AI crowd may be set out in detail as follows:

15. Baker, "Why Computer's Can't Act," p. 159.
16. Ibid., p. 160.
17. Ibid.
18. Ibid.

C	1.	Necessarily, X is an original agent only if X has the capacity to formulate intentions.	
	2.	Necessarily, X has the capacity to formulate intentions only if X has an irreducible first person perspective.	
	3.	Necessarily, X has an irreducible first person perspective only if X has second-order consciousness.	
	4.	Necessarily, X has second-order consciousness only if X has self-consciousness.	
∴	5.	Necessarily, X is an original agent only if X has self-consciousness	1,2,3&4
	6.	Necessarily, X is a machine only if X is designed and programmed.	
	7.	Necessarily, X is designed and programmed only if X operates just according to rule-governed transformations on discrete input.	
	8.	Necessarily, X operates just according to rule-governed transformations on discrete input only if X lacks self-consciousness.	
∴	9.	Necessarily, X is a machine only if X lacks self-consciousness.	6,7&8
∴	10.	Necessarily, X is a machine only if X is not an original agent.	5&9

A Measured Response on Behalf of the AI Crowd

While there presumably exist skeptical challenges which ought not be taken seriously because they are, for want of careful argumentation, themselves unserious, I submit that Baker's skeptical challenge to the AI crowd is serious and ought to be taken as such. It calls for a measured response. It would be a mistake, in other words, for the AI crowd to dismiss Baker's challenge out of hand for want of technical sophistication, say, in the absence of decisive counterarguments. Moreover, counterarguments will not be decisive if they simply ignore the underlying import of the skeptic's claims.

For example, given the weight of argument against physicalist solutions to the hard problem of consciousness generally, it would be incautious of the AI crowd to respond by rejecting C.8 (but see[19] for a comprehensive review of the hard problem). In simple terms, the AI crowd should join the mind crowd in finding it daft at this point for a roboticist to claim that *there is something it is like to be her robot*, however impressive the robot or resourceful the roboticist in building it.

A more modest strategy is to sidestep the hard problem of consciousness altogether by arguing that having an irreducible FPP is not, contrary to C.2, a necessary condition

19. D. Chalmers, "Consciousness and Its Place in Nature," Philosophy of Mind: Classical and Contemporary Readings (Oxford: Oxford University Press, 2002), 247–272.

on the capacity to form intentions. This is the appropriate point to press provided that it also appeals to the mind crowd's own concerns. For instance, if it can be argued that the requirement of an irreducible FPP is too onerous even for persons to formulate intentions under ordinary circumstances, then Baker's assumption of Castaneda's account will be vulnerable to criticism from both sides. Working from the other direction, it must also be argued the notion of programming that justifies C.7 and C.8 is far too narrow even if we grant that programming an irreducible FPP is beyond our present abilities. The measured response I am presenting thus seeks to moderate the mind crowd's excessively demanding conception of intention while expanding their conception of programming so as to reconcile, in principle, the *prima facie* absurdity of a programmed (machine) intention.

Baker's proposal that the ability to form intentions implies an irreducible FPP is driven by her adoption of Castaneda's[20] analysis of intention: To formulate an intention to A is to endorsingly think the thought, "I shall do A". There are, however, other analyses of intention which avoid the requirement of an irreducible FPP. Davidson[21] sketches an analysis of what it is to form an intention to act: "an action is performed with a certain intention if it is caused in the right way by attitudes and beliefs that rationalize it."[22] Thus,

> If someone performs an action of type A with the intention of performing an action of type B, then he must have a pro-attitude toward actions of type B (which may be expressed in the form: an action of type B is good (or has some other positive attribute)) and a belief that in performing an action of type A he will be (or probably will be) performing an action of type B (the belief may be expressed in the obvious way). The expressions of the belief and desire entail that actions of type A are, or probably will be, good (or desirable, just, dutiful, etc.).[23]

Davidson is proposing that S A's with the intention of B-ing only if

i. S has pro-attitudes towards actions of type B.

ii. S believes that by A-ing S will thereby B.

The pro-attitudes and beliefs S has which rationalize his action cause his action. But, of course, it is not the case that S's having pro-attitudes towards actions of type B and S's believing that by A-ing she will thereby B jointly implies that S actually A's with the intention of B-ing. (i) and (ii), in simpler terms, do not jointly suffice for S's A-ing with the intention of B-ing since it must be that S A's because of her pro-attitudes and beliefs. For Davidson, 'because' should be read in its causal sense. Reasons consisting as they do of pro-attitudes and beliefs cause the actions they rationalize.

20. Castaneda, *Thinking and Doing: The Philosophical Foundations of Institutions*.
21. D. Davidson, "Intending," Essays on Actions and Events (Oxford: Clarendon Press, 1980), 83–102.
22. Ibid., p. 87.
23. Ibid., pp. 86-87.

Causation alone is not enough, however. To suffice for intentional action reasons must cause the action in the right way. Suppose (cf[24]) Smith gets on the plane marked 'London' with the intention of flying to London, England. Without alarm and without Smith's knowledge, a shy hijacker diverts the plane from its London, Ontario destination to London, England. Smith's beliefs and pro-attitudes caused him to get on the plane marked 'London' so as to fly to London, England. Smith's intention is satisfied, but only by accident, as it were. So it must be that Smith's reasons cause his action in the right way, thereby avoiding so called wayward causal chains. Hence, S A's with the intention of B-ing if, and only if,

i. S has pro-attitudes towards actions of type B.

ii. S believes that by A-ing S will thereby B.

iii. S's relevant pro-attitudes and beliefs cause her A-ing with the intention of B-ing in the right way.

Notice that there is no reference whatsoever involving an irreducible FPP in Davidson's account. Unlike Castaneda's account, there is no explicit mention of the first person indexical. So were it the case that Davidson thought animals could have beliefs, which he does not,[25] it would be appropriate to conclude from Davidson's account that animals can act intentionally despite worries that animals would lack an irreducible first-person perspective. Presumably robots would not be far behind.

It is nevertheless open to Baker to ask about (ii): S believes that by A-ing S will thereby B. Even if S does not have to explicitly and endorsingly think, "I shall do A" to A intentionally, (ii) requires that S has a self-referential belief that by A-ing he himself will thereby B. Baker can gain purchase on the problem by pointing out that such a belief presupposes self-consciousness every bit as irreducible as the FPP.

Consider, however, that a necessary condition on Davidson's account of intentional action is that S believes that by A-ing S will thereby B. Must we take 'S' in S's belief that by A-ing S will thereby B *de dicto*? Just as well, could it not be the case (*de re*) that S believes, of itself, that by A-ing it will thereby B?

The difference is important. Taken *de dicto*, S's belief presupposes self-consciousness since S's belief is equivalent to having the belief, "by A-ing I will thereby B". Taken (*de re*), however, S's belief presupposes at most self-*representation*, which can be tokened without solving the problem of (self) consciousness.

Indeed, it does not seem to be the case that the intentions I form presuppose either endorsingly thinking "I shall do A!" as Castaneda (and Baker) would have it or a *de dicto* belief that by A-ing I will B as Davidson would have it. Intention-formation is transparent: I simply believe that A-ing B's, so I A. The insertion of self-consciousness as an intermediary requirement in intention formation would effectively eliminate many intentions in light of environmental pressures to act quickly. Were Thog the caveman required to endorsingly think "I shall climb this tree to avoid the saber-toothed tiger" before scrambling up the tree he would lose precious seconds

24. Davidson, "Intending," pp. 84-85.

25. D. Davidson, "Thought and Talk," Inquiries into Truth and Interpretation (Oxford: Clarendon Press, 1984), 155–170.

and, very likely, his life. Complexity, particularly temporal complexity, constrains us as much as it does any putative original machine agent. A theory of intention which avoids this trouble surely has the advantage over theories of intention which do not.

In a subsequent pair of papers[26] and a book,[27] Baker herself makes the move recommended above by distinguishing between weak and strong first-person phenomena (later recast in more developmentally discerning terms as 'rudimentary' and 'robust' first-person perspectives), on the one hand, and between minimal, rational, and moral agency, on the other. Attending to the literature in developmental psychology (much as many in the AI crowd have done and would advise doing), Baker[28] argues that the rudimentary FPP is properly associated with minimal–that is, non-reflective–agency, which in turn is characteristic of infants and pre-linguistic children and adult animals of other species. Notably, the rudimentary FPP does *not* presuppose an *irreducible* FFP, although the robust FPP constituitively unique to persons does. As Baker puts it,

> [P]ractical reasoning is always first personal: The agent reasons about what to do on the basis of her own first-person point of view. It is the agent's first-person point of view that connects her reasoning to what she actually does. Nevertheless, the agent need not have any first-person concept of herself. A dog, say, reasons about her environment from her own point of view. She is at the origin of what she can reason about. She buries a bone at a certain location and later digs it up. Although we do not know exactly what it's like to be a dog, we can approximate the dog's practical reasoning from the dog's point of view: Want bone; bone is buried over there; so, dig over there. The dog is automatically (so to speak) at the center of the her world without needing self-understanding.[29]

Baker further argues in these pages[30] that, despite the fact that artifacts like robots are intentionally made for some purpose or other while natural objects sport no such teleological origin, "this differences does not signal any ontological deficiency in artifacts *qua* artifacts". Artifacts suffer no demotion of ontological status insofar as they are ordinary objects regardless of origin. Her argument, supplemented and supported by Amie L. Thomasson,[31] repudiates drawing on the distinction between mind-dependence and mind-independence (partly) in light of the fact that,

> [A]dvances in technology have blurred the difference between natural objects and artifacts. For example, so-called digital organisms are computer programs that (like biological organisms) can mutate, reproduce, and compete with one another. Or consider robo-ratsrats with implanted electrodes

26. L.R. Baker, "The First-Person Perspective: A Test for Naturalism," *American Philosophical Quarterly* 35, no. 4 (1998): 327–348; L.R. Baker, "First-Personal Aspects of Agency," *Metaphilosophy* 42, nos. 1-2 (2011): 1–16.

27. L.R. Baker, *Naturalism and the First-Person Perspective* (New York: Oxford University Press, 2013).

28. Baker, "First-Personal Aspects of Agency."

29. Baker, *Naturalism and the First-Person Perspective*, p. 189.

30. L.R. Baker, "The Shrinking Difference Between Artifacts and Natural Objects," *APA Newsletter on Philosophy and Computers* 07, no. 2 (2008): 2–5.

31. A.L. Thomasson, "Artifacts and Mind-Independence: Comments on Lynne Rudder Baker's 'The Shrinking Difference between Artifacts and Natural Objects'," *APA Newsletter on Philosophy and Computers* 08, no. 1 (2008): 25–26.

> that direct the rats movements. Or, for another example, consider what one researcher calls a bacterial battery : these are biofuel cells that use microbes to convert organic matter into electricity. Bacterial batteries are the result of a recent discovery of a micro-organism that feeds on sugar and converts it to a stream of electricity. This leads to a stable source of low power that can be used to run sensors of household devices. Finally, scientists are genetically engineering viruses that selectively infect and kill cancer cells and leave healthy cells alone. Scientific American referred to these viruses as search-and-destroy missiles. Are these objects-the digital organisms, robo-rats, bacterial batteries, genetically engineered viral search-and-destroy missilesartifacts or natural objects? Does it matter? I suspect that the distinction between artifacts and natural objects will become increasingly fuzzy; and, as it does, the worries about the mind-independent/mind-dependent distinction will fade away.[32]

Baker's distinction between rudimentary and robust FPPs, suitably extended to artifacts, may cede just enough ground to the AI crowd to give them purchase on at least *minimal* machine agency, all while building insurmountable ramparts against the AI crowd to defend, on behalf of the mind crowd, the special status of persons, enjoying as they must their computationally intractable robust FPPs. Unfortunately Baker does not explain precisely how the minimal agent enjoying a rudimentary FPP develops into a moral agent having the requisite robust FPP. That is, growing children readily, gracefully, and easily scale the ramparts simply in the course of their normal development, yet how remains a mystery.

At most we can say that there are many things a minimal agent cannot do rational (reflective) and moral (responsible) agents can do. Moreover, the mind crowd may object that Baker has in fact ceded no ground whatsoever, since even a suitably attenuated conception of intention cannot be programmed under Baker's conception of programming. What is her conception of programming? Recall that Baker defends B.3 by arguing that machines cannot achieve a first-person perspective since machines gain information *only* through rule-based transformations on discrete input and no amount or combination of such transformations could suffice for the transition from a third-person perspective to a first-person perspective. That is,

32. Baker, "The Shrinking Difference Between Artifacts and Natural Objects," p. 4.

D	1.	If machines were able to have a FPP, then the FPP can be the result of transformations on discrete input via specifiable rules.	
	2.	If the FPP can be the result of transformations on discrete input via specifiable rules, then there exists some amount of third-person information which compels a shift to first-person knowledge.	
	3.	No amount of third-person information compels a shift to first-person knowledge.	
∴	4.	First-person episodes cannot be the result of transformations on discrete input via specifiable rules.	2&3
∴	5.	Machines necessarily lack an irreducible first-person perspective.	1&4

The problem with D is that it betrays an overly narrow conception of machines and programming, and this is true even if we grant that we don't presently know of *any* programming strategy that would bring about an irreducible FPP.

Here is a simple way of thinking about machines and programming as Argument D would have it. There was at one time (for all I know, there may still be) a child's toy which was essentially a wind-up car. The car came with a series of small plastic disks, with notches around the circumference, which could be fitted over a rotating spindle in the middle of the car. The disks acted as a cam, actuating a lever which turned the wheels when the lever hit a notch in the side of the disk. Each disk had a distinct pattern of notches and resulted in a distinct route. Thus, placing a particular disk on the car's spindle 'programs' the car to follow a particular route.

Insofar as it requires that programming be restricted to transformations on discrete input via specifiable rules, Argument D treats all machines as strictly analogous to the toy car and programming as analogous to carving out new notches on a disk used in the toy car. Certainly Argument D allows for machines which are much more complicated than the toy car, but the basic relationship between program and machine behavior is the same throughout. The program determines the machine's behavior, while the program itself is in turn determined by the programmer. It is the point of D.2 that, if an irreducible FPP were programmable, it would have to be because the third-person information which can be supplied by the programmer suffices for a first-person perspective, since all the machine has access to is what can be supplied by a programmer.

Why should we think that a machine's only source of information is what the programmer provides? Here are a few reasons to think that machines are not so restricted:

- Given appropriate sensory modalities and appropriate recognition routines, machines are able to gain information about their environment without that information having been programmed in advance.[33] It would be as if the toy car had an echo-locator on the front and a controlling disk which notched itself in reaction to obstacles so as to maneuver around them.

33. R.C. Arkin, *Behavior Based Robotics* (Cambridge, Mass.: MIT Press, 1998).

- Machines can be so constructed as to 'learn' by a variety of techniques.[34] Even classical conditioning techniques have been used. The point is merely that suitably constructed, a machine can put together information about its environment and itself which is not coded in advance by the programmer and which is not available other than by, for example, trial and error. It would be as if the toy car had a navigation goal and could adjust the notches in its disk according to whether it is closer or farther from its goal.

- Machines can evolve.[35] Programs evolve through a process of mutation and extinction. Code in the form of so-called genetic algorithms is replicated and mutated. Unsuccessful mutations are culled, while successful algorithms are used as the basis for the next generation. Using this method one can develop a program for performing a particular task without having any knowledge of how the program goes about performing the task. Strictly speaking, there is no programmer for such programs. Here the analogy with the toy car breaks down somewhat. It's as if the toy car started out with a series of disks of differing notch configurations and the car can take a disk and either throw it out or use it as a template for further disks, depending on whether or not a given disk results in the car being stuck against an obstacle, for instance.

- Programs can be written which write their own programs.[36] A program can spawn an indefinite number of programs, including an exact copy of itself. It need not be the case that the programmer be able to predict what future code will be generated, since that code may be partially the result of information the machine gathers, via sensory modalities, from its environment. So, again, in a real sense there is no programmer for these programs. The toy car in this case starts out with a disk which itself generates disks and these disks may incorporate information about obstacles and pathways.

Indeed, many of the above techniques develop Turing's own suggestions:

> Let us return for a moment to Lady Lovelace's objection, which stated that the machine can only do what we tell it to do...
>
> Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

34. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 3rd (Cambridge, Mass.: MIT Press. A Bradford Book, 1998).
35. D. H. Ballard, *An Introduction to Natural Computation* (Cambridge, Mass.: MIT Press, 1997).
36. Ibid.

> We have thus divided our problem into two parts. The child programme and the education process. These two remain very closely connected. We cannot expect to find a good child machine at the first attempt. One must experiment with teaching one such machine and see how well it learns...
>
> The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. The reader may draw a parallel with the Constitution of the United States.[37]

As Turing anticipated, machines can have access to information and utilize it in ways which are completely beyond the purview of the programmer. So while it may not be the case that a programmer can write code for an irreducible FPP, as Argument D requires, it still can be argued that the sources of information available to a suitably programmed robot nevertheless enable it to formulate intentions when intentions do not also presuppose an irreducible FPP.

Consider the spectacularly successful Mars rovers Spirit and Opportunity. Although the larger goal of moving from one location to another was provided by mission control, specific routes were determined *in situ* by constructing maps and evaluating plausible routes according to obstacles, inclines, etc. Thus the Mars rovers were, in a rudimentary sense, gleaning information from their environment and using that information to assess alternatives so as to plan and execute subsequent actions. None of this was done with the requirement of, or pretense to having, an irreducible FPP, yet it does come closer to fitting the Davidsonian model of intentions. To be sure, this is intention-formation of the crudest sort, and it requires further argument that propositional attitudes themselves are computationally tractable.

A Larger Point: Avoiding Excesses on Both Sides

Baker closes her original article by pointing out that robots' putative inability to form intentions has far-reaching implications:

> So machines cannot engage in intentional behavior of any kind. For example, they cannot tell lies, since lying involves the intent to deceive; they cannot try to avoid mistakes, since trying to avoid mistakes entails intending to conform to some normative rule. They cannot be malevolent, since having no intentions at all, they can hardly have wicked intentions. And, most significantly, computers cannot use language to make assertions, ask questions, or make promises, etc., since speech acts are but a species of intentional action. Thus, we may conclude that a computer can never have a will of its own.[38]

37. A.M. Turing, "Computing Machinery and Intelligence," *Mind* 59 (1950): pp. 454-458.
38. Baker, "Why Computer's Can't Act," p. 163.

The challenge for the AI crowd, then, is to break the link Baker insists exists between intention formation and an irreducible FPP in its robust incarnation. For if Baker is correct and the robust FPP presupposes self-consciousness, the only way the roboticist can secure machine agency is by solving the vastly more difficult problem of consciousness, which so far as we presently know is a computationally impenetrable problem. I have argued that the link can be broken, provided a defensible and computationally tractable account of intention is available to replace Castaneda's overly demanding account.

If my analysis is sound, then there are times when it is appropriate for the AI crowd to push back against the mind crowd. Yet they must do so in such a way as to respect so far as possible the ordinary notions the mind crowd expects to see employed. In this case, were the AI crowd to so distort the concept of intention in their use of the term that it no longer meets the mind crowd's best expectations, the AI crowd would merely have supplied the mind crowd with further skeptical arguments. In this sense, the mind crowd plays a valuable role in demanding that the AI crowd ground their efforts in justifiable conceptual requirements, which in no way entails that the AI crowd need accept those conceptual requirements without further argument. Thus the enterprise of artificial intelligence has as much to do with illuminating the efforts of the philosophers of mind as the latter have in informing those working in artificial intelligence.

This is a plea by example, then, to the AI crowd that they avoid being overly satisfied with themselves simply for simulating interesting behaviors, unless of course the point of the simulation simply is the behavior. At the same time, it is a plea to the mind crowd that they recognize when their claims go too far even for human agents and realize that the AI crowd is constantly adding to their repertoire techniques which can and should inform efforts in the philosophy of mind.

References

Arkin, R.C. *Behavior Based Robotics*. Cambridge, Mass.: MIT Press, 1998.

Baker, L.R. "First-Personal Aspects of Agency." *Metaphilosophy* 42, nos. 1-2 (2011): 1–16.

———. *Naturalism and the First-Person Perspective*. New York: Oxford University Press, 2013.

———. "The First-Person Perspective: A Test for Naturalism." *American Philosophical Quarterly* 35, no. 4 (1998): 327–348.

———. "The Shrinking Difference Between Artifacts and Natural Objects." *APA Newsletter on Philosophy and Computers* 07, no. 2 (2008): 2–5.

———. "Why Computer's Can't Act." *American Philosophical Quarterly* 18 (1981): 157–163.

Ballard, D. H. *An Introduction to Natural Computation*. Cambridge, Mass.: MIT Press, 1997.

Castaneda, H-N. *Thinking and Doing: The Philosophical Foundations of Institutions*. Dordrecht: D. Reidel Publishing Co., 1975.

Chalmers, D. "Consciousness and Its Place in Nature," 247–272. Philosophy of Mind: Classical and Contemporary Readings. Oxford: Oxford University Press, 2002.

Davidson, D. "Intending," 83–102. Essays on Actions and Events. Oxford: Clarendon Press, 1980.

———. "Thought and Talk," 155–170. Inquiries into Truth and Interpretation. Oxford: Clarendon Press, 1984.

Sutton, R. S., and A. G. Barto. *Reinforcement Learning: An Introduction*. 3rd. Cambridge, Mass.: MIT Press. A Bradford Book, 1998.

Thomasson, A.L. "Artifacts and Mind-Independence: Comments on Lynne Rudder Baker's 'The Shrinking Difference between Artifacts and Natural Objects'." *APA Newsletter on Philosophy and Computers* 08, no. 1 (2008): 25–26.

Turing, A.M. "Computing Machinery and Intelligence." *Mind* 59 (1950): 433–60.