A Theologian Looks at AI

Andrew P. Porter
Center for Theology and the Natural Sciences,
Graduate Theological Union
Berkeley, CA

Abstract

AI has a long history of making fine tools, and an equally long history of trying to simulate human intelligence, without, I contend, really understanding what intelligence consists in: the ability to deal with the world, which presupposes having a stake in one's own being. The tools are very nifty, but I don't see how it's even possible to simulate having a stake in one's own being (Heidegger, Kierkegaard).

The prospectus for the Conference includes "theological perspectives on human nature and the potential impacts that AI-related sciences and technologies incur in current and future society." My background: just a scruffy retread from physics having fun in theology and hermeneutics. My perspective on AI is that of an outsider, one trained in computational physics (which is largely irrelevant) and also in theological hermeneutics, which might matter. What follows is mostly questions, with some context.

There is a distinction between the sort of being of tools and that of living organisms (especially humans, language-capable life). They be what they are in quite different ways. And while it appears from a distance that AI has been producing tools, some very nifty tools, the idea of producing artificial "intelligence" puzzles me. I don't understand what *intelligence* means in this context. It doesn't mean just the ability to do arithmetic; if that's all intelligence is, then artificial intelligence is several millenia old: abaci have artificial intelligence, and recent AI is just producing more of the same, only magnificently bigger and better.

A colleague once opined that Luddites are wrong to be spooked by AI: Two hundred years ago, steam engines proved able to do what men with shovels or horses pulling vehicles do, only better, and on a much bigger scale. By the same token, computers do some of what humans have long done, but faster, more accurately, and on a much bigger scale.

Another colleague, Cindy Mason, once opined about the difference between a robot or AI and a human: "It wouldn't have my problems; or personality, humor, a history, empathy."[1] So it is not obvious what is happening in AI and cognitive science.

[1]Private communication, 1984 January 27.

My bewilderment runs deeper. Do AI researchers think they are making more tools, or simulating "intelligence"? What do they think "intelligence" *is*? The little bits of conversation that I have overheard don't answer that last question. What's the difference, for AI, between intelligence and tools? If there is one?

More generally, what kinds of beings is AI trying to create? And if it is trying to simulate one or another kind of being, what mode of being does the simulation or representation have? And what is the difference between a simulation and a representation? How are they different from the thing itself, as it is in the wild? And for whom might a representation be a representation?

There are answers to some of these questions from the Continental tradition in philosophy, and the answers can stand as probes to the thinking of AI. Conjecture: "intelligence" in the discourse of AI is a place-holder for the features and capacities of human beings (language-capable life). If the term "intelligence" goes unexamined, a lot gets missed with it.

Because I am a transient visitor to AI, let me, in order to keep these remarks manageable, take as point of departure an article (Dreyfus 2007) in which Hubert Dreyfus reviewed (Wheeler 2005). In support of his critique of Wheeler, Dreyfus drew heavily on (Freeman 2000) and (Freeman 1995). My own comments will extend Dreyfus's critique and base on it some questions for AI.

The thrust of Dreyfus's critique of Wheeler and others in AI is that their programs do not do what humans do under the aspect of Being-in-the-World. Those arguments about Heideggerian AI can mostly be left to their participants without trying to summarize or evaluate them. I know a little about Heidegger but not much about AI. It is fair to observe that neither Dreyfus nor any that I could find quickly probed beneath Being-in-the-World to the structure of Dasein as the sort of being that has a stake or interest in its own being.

Note especially that Being-in-the-World is here related to and (as the argument plays out in *Being and Time*) grounded in the structure of Dasein as the sort of being that has a stake in its own being. This is announced early (German p. 12) but developed only slowly (after Being-in-the-World), as Care, anxiety, mortality, and temporality. Dasein has stakes or interests in both itself and in things in the world. Care and the other deeper levels of the Daseinanalytik appear in Dreyfus's

review of Wheeler under the words "cope" and "coper," one who copes. The word "cope" very nicely bypasses all the many problems in the deeper structure of Dasein.

Dreyfus notes more than once that world is not the same thing as the universe: the universe is just the collection of all beings. World enables humans to discriminate between things in the world.[2] World and worldhood are human-relative. The universe, obviously, is not. The world enables humans to discriminate between figure and background, between the relevant and the irrelevant. These are both variants of distinguishing between what *matters* and what doesn't. AI in general, and even Heideggerian AI, apparently have difficulty with these distinctions.

Beneath the ability to discriminate things from background lies the problem of what holds a thing together as *one* thing, and not just a heap of molecules. This used to be called formal causes, an Aristotelian concept, but formal causes (substantial forms) did something horrible to offend at the turn of the seventeenth century, and there is lasting bitterness against formal causes still today. It would be better just to call the thing of interest "ontological glue," what holds a thing together. That term names a problem, not its solutions. There are many kinds of ontological glue, just as there are many modes of being. Even the term "glue" may be risky. It makes it sound as if glue is an occult kind of matter, holding ordinary matter together in "things." To deny that (as I do) is to take sides in Scholastic controversies that do not belong in this paper. Better would be that glue is not a thing, it is the way matter holds together in things that are made of matter. My solution is not always Aristotelian, and the what-it-is of a thing is not always inherent in the bearer. Heidegger showed that: a chair is a chair not because it has some particular form but because it is equipment for sitting *by humans*.

Being-in-the-World and interpreting the world, the locus of research and controversies in AI, came first in Division I of *Being and Time*. What came next, though it was announced at the beginning, lay beneath and more primordially than Being-in-the-World. It is Dasein's structure as having a stake in its own being, developed in the phenomenology of care and temporality. Here is Heidegger's definition: "Dasein is that entity which, as Being-in-the-World, is an issue for itself." [3] To say that Dasein "is an issue for itself" is to say that Dasein has a stake in its own being, that Dasein has interests in its own being. We shall modify this definition shortly, but this is where Heidegger started.

There is another formulation of the idea of mattering to oneself, older than Heidegger, in Kierkegaard's *Sickness Unto Death*. In paraphrase, a human self is "a relationship that relates itself to itself" ((Kierkegaard 1941), Lowrie trans., p. 162). but is constituted as such by an Other. I would like to set aside the leap of logic to an Other and note that a human self, or Heidegger's Dasein, is first constituted as such by many *others*. Contrary to the silence in Heidegger's definition, it is not just the Dasein in focal view that has an interest in its own being; many other Daseins do also, and the Dasein in view has reciprocal interests in their being.[4] (This applies to life more generally than just human beings; a member of one species cannot be what it is without other members of the same species.) The character and extent of a Dasein's interests in other Daseins probably cannot be exhaustively enumerated. This is incidentally one reason why a Dasein cannot be construed as a system that has a state: determinants of its being are spread out too elusively in the world, far beyond that particular Dasein itself, both in space *and in time*.

Even though a Dasein is not a system, its focal material substrate (colloquially its "body") *is* a system. Since a Dasein is not a system, it does not have a state, whatever may be said of its body, including its brain. The same goes for human actions. It used to be thought that action was a species of caused motion, but that model fails in many places. Instead, action is the result of a narrative selection of motions, but the relation of narrative and action is circular.[5] It is impossible to know what motions to include as relevant without knowing a narrative, and impossible to narrate without knowing something of the possible motions one might include. Narrative is possible wherever contingency affects someone's interests. Actions are constituted as what they are by much that is beyond their substrates, and so they cannot be systems or changes of states of systems.

The claim that Daseins are constituted by others can be corroborated (at the level of existentiell modifications) in the sociology of knowledge, from the structure of primary socialization.[6] Primary socialization can happen in many ways, but the *capacity* for primary socialization is part of Dasein's ontology, part of its mode of being. Human being is the sort of being capable of this kind of existential interaction. Without that capacity, something could not even be a failed human being.

There are many related definitions of Dasein, but to save ink (you are allowed to laugh), let us use Kierkegaard's language and call a Dasein a *r*elationship that *r*elates *i*tself *t*o *i*tself — *c*onstituted *a*s *s*uch *b*y *o*thers *l*ike *i*t: a RRITICASBOLI.

Ontological involvements with other people bring many things. One of them is the capacity for emotions. Cindy Mason, among other AI researchers, has been puzzled and dissatisfied by the lack of emotion in robots and AI, and surmised that AI is unlikely to succeed without in some way having or getting emotions (Mason 2008), (Damasio 1995). Another thing that comes with interinvolvement is identification with significant others, without which primary socialization is impossible. Identification is about ontology, even if it is poorly explored. I *am* like this or that other, and

[2] This distinction appeared originally in (Heidegger 1960) section 14, German p. 64–65; p. 93 of the Macquarrie and Robinson translation.

[3] (Heidegger 1960), p. 142, German pagination. It first appears on page 12 and again on 191; maybe more. The German is colloquial, and both English translations are clearer.

[4] See (Porter 2011), sections 3.4.1 and 3.4.2.

[5] (Porter 2011), section 4.5.2, p. 120, and passim (see the index, for circularity).

[6] (Berger and Luckmann 1966), section III.1.a, "Primary Socialization."

the likeness is part of my being, not just something incidental. Human beings are capable of this, tools are not, plants probably are not, other animals are capable only to a limited extent. It is language that gives human involvements with others and the world their enormous reach.

My central question is, How would an AI program propose to turn universe into world and discriminate things within the world without in some sense being a RRITI-CASBOLI? The relationship in a RRITI is always physical and embodied, and it consists of Dasein's interests and involvements, advantages and disadvantages, in projection of possible futures — its own and the world's. AI (so far, and to the best of my limited knowledge) has not even attempted to be, model, represent, or simulate a RRITI-CASBOLI. There does not appear to be much attention to the mode of being of the artifacts of AI.

What AI does attempt: Representation, modeling, simulation — but of a very different sort of mind. Dreyfus notes the sources of AI in early modern philosophers.[7] AI is the continuation of a tradition that goes back to the beginning, still seeking "what Plato sought: a world in which the possibility of clarity, certainty, and control is guaranteed; a world of data structures, decision theory, and automation."[8] AI would model humans, and calculate narratives, ignoring the openness and ambiguity of narratives.[9] AI would like to solve the "common sense" problem, i. e., model human knowledge of the everyday world. That is what Heideggerian phenomenology calls "Being-in-the-World," understanding, and interpretation, in Sections 31–34 of *Being and Time*.

What AI does *not* yet attempt: to construct something computational that could know itself, have emotions, relate to itself or to others, relate to its past and its future, have a stake in its own being. My contention is that only a RRITI-CASBOLI can have a world and discriminate things in the world. To say that something in the world *matters* is to say that it matters *for someone*; i. e., for a RRITI-CASBOLI. Is AI trying to make a computational RRITI-CASBOLI?

In Dreyfus's account of Walter Freeman's model, it would appear that the relation of rabbit and world, proposed as model for AI solving the "what matters" problem, works by coupling the rabbit and the world — two systems with interactions. This is an analogy from thermodynamics, though Dreyfus did not remark that.[10] Thinking of the universe as coupled systems was latent in Newtonian physics, but it came into its own with nineteenth-century thermodynamics and statistical mechanics. The universe gets subdivided into systems and subsystems, and they interact. Both matter and energy cross the boundaries of a system, and everything is comprehensible (indeed, calculable to high accuracy) in terms of systems that have states that change as functions of time. Therein lies the coupling between systems. But this is just an analogy, and there are many things about beings that matter to themselves that it can't really account for. Dreyfus and Freeman could I suppose cite IV Lateran, "One may not note similarities between creator and creatures without also noting greater dissimilarities."[11] but I don't think that would cut much ice.

It is hard to understand how it might be possible to simulate mattering without thereby being the sort of being for which things matter. What would it mean for AI to model a RRITI-CASBOLI? I have no idea.

AI might very well reach similar results by another route — after all, dictionaries and (in mathematical physics) computers do just that. The AI ontology projects,[12] collecting conceptual frames, are, in effect a dictionary with the capacity to calculate links between dictionary entries and resolve at least some of the problems that arise between entries. Dictionaries are sometimes right, and so are quite useful. Insofar as the distributed ontologies project in AI (not quite the same thing as the distributed ontology in *Living in Spin*) maps the principal frames[13] in a language, it will produce a very useful tool. But to be useful is to be a tool, not a living organism, a Dasein, or a RRITI-CASBOLI. One index of the limits of dictionaries and of frames more generally is haiku. The goal in a haiku is to pack as much world into as few words as possible, usually going far beyond meanings that could appear in any dictionary.[14] Another index of the limits of dictionaries is the failure of natural language translation strategies that rely on one-to-one correspondences between dictionaries. One such rendered "out of sight, out of mind" as "invisible, insane."

So there are many questions for AI. What is AI trying to do? Model humans? That is called philosophical anthropology. If it is central to a basic life orientation (BLO), it is a theological anthropology. One may further ask, theology of what BLO? One possibility is the program of what goes by the name of naturalism, usually accompanied by some form of nominalism and the materialism that ignores the problem of formal causes, ontological glue. Since the seventeenth century, natural laws have appeared to many to be the only way to construe the world and ultimate reality as orderly and reliable after the loss of the medieval cosmic synthesis (Dupré 1993). Violations of natural laws were forbidden both to safeguard that reliability and also to protect the integrity of naturalistic questioning. There is no other way to do science, but more than science was involved: the comforts of a BLO. There are many things that transcend natural laws without in any way violating natural laws: narrative (and so also history), human action, phenomenology, hermeneutics, beauty, unanswerable questions — or just Peter Berger's list: order, play, hope, humor, and grace (Berger 1969). As going beyond scientific explanation, these things

[7] In the first paragraph of "Why Heideggerian AI Failed."

[8] (Dreyfus 1972), p. 212.

[9] (Porter 2011), chapters 3–5.

[10] This was not the only uncriticized analogy, but most of them can be recognized in contrast to the background of the distributed ontology of human action in *Living in Spin*.

[11] After IV Lateran (1215 CE): http://www.fordham.edu/ halsall/basis/lateran4.html.

[12] These are new to me. The Wiki article, "Ontology (information science)" might work as a starting point.

[13] In the sense of (Lakoff 1987) not (McCarthy and Hayes 1969).

[14] See for example (Ball 2012).

have the potential to disturb the comfort of naturalistic order. The easy remedy is to pretend that natural laws will soon explain everything, everything at all. The classic elaboration of the differences is in (Eliade 1959), where pre-scientific naturalistic BLO is explored in detail. You can live that way if you like, but we are not required to keep a straight face.

More questions arise: What is the ontological status of models, simulations, representation, or naratives? With respect to systems of partial differential equations (as in fluid mechanics) or of ordinaries (as in celestial mechanics), the ontological status of models and simulations is not problematic. But is AI trying to do by other means what humans *do* or model what humans *be*? How is AI to handle the ontological objection to its models that "the map is not the territory"? In computational fluid mechanics, the whole point of the simulation (the map) is the get around in the territory (build working fluid mechanical hardware). It is not so obvious what AI is trying to do.

AI has produced computers that have hands that "handle," eyes that "see," ears that "hear," tongues that "speak." But do the makers of AI really want to become like their artifacts?

One may ask, what could you do with a model of human being? One obvious use is to get conceptual control over human being (and practically, over human beings). This doubtless opens the door to abuses of other people.

What would getting conceptual control over human being mean? To be human would in some sense then mean to be something that we have made, ourselves, rather than being something simply given to us by whatever you want credit in your narrative (evolution, for one obvious example). AI's quest for control is a quest for self-mastery.

Assimilating organisms to artifacts is not a new idea in the modern world. It goes back at least to Paley, and it is the central hidden mistake in biblical Creationism. Is this what AI wants to do? Assimilating organisms and artifacts is implicit in using computations (computational tools, zuhanden) to model Dasein — which has a completely different mode of being. Creationism would treat humans as artifacts in order to evade the challenges of living with critical history; AI would treat humans as artifacts in order to control the challenges of being human. AI *as anthropology* is the creationism of a certain kind of geek subculture.

What if AI as quest for simulation and control fails? There are at least three possible responses: (1) to persist in seeking control; (2) to give up philosophical anthropology entirely; (3) to accept being the beings we are given to be — included in which is a challenge to do something with what we are given to be. The second may be passed over, for though one may give up this route to a philosophical anthropology, there are ways other than the Platonist tradition culminating in AI to get an anthropology. (1) is ambiguous. As AI, it has (so far) failed, and there is no prospect of success soon, inasmuch as AI doesn't even have a plan to simulate a RRITI-CASBOLI (though successful abuse of others is entirely possible). If it persists (and still fails), it is likely to manifest tragedy (as in waste of time and effort — or in abuses). If it acknowledges its failure, it becomes (3) — like Jacob at the Jabbok: Israel, or he who struggles with ultimate reality. (3) leads back to other means of philosophical anthropology, as in phenomenology.

There is another choice for AI/neurophysiology, falling within (3), and I think it has been well under way for some time now. Examine the neurophysiological substrates in human actions when the human does such-and-such. This presupposes prior knowledge of what the human subject of such an inquiry is doing in "such-and-such." That knowledge is not naturalistic but phenomenological, a kind of narrative, and the relation to the naturalistic substrate is diagnostic, not something reducible to naturalistic categories. This has been explored in (Ricoeur 1966) and (Reagan 1968). Ricoeur argued for a disentanglement of existential and naturalistic explanatory categories. Because such a diagnostic relation presupposes a *prior* knowledge of what humans are and do it cannot be used to *explain* what humans are and do. In effect, material cause (what the substrate of a thing is) is not a substitute for formal cause (the ontological glue that holds the thing together and explains its mode of being).

Because the material causes can always be taken as an explanation, AI is then ambiguous: It can be many things, at once, or even become something new after the fact. This is part of the ambiguity of narrative, which has metastasized to human action, because action is a circular synthesis of motions and narrative. In the end we come to the radical ambiguity of AI: It is not clear what it is doing. In some cases, this arises from inattention to what it is doing, but there is a deeper source of ambiguity simply because AI research itself is a collection of human actions. What an act is depends on how it is narrated, and that is open. What an act is depends on its consequences, foreseen or not, desired or not. When desired consequences are not spelled out, there may be trouble ahead.[15] Above all, the question, "what does AI think it is doing?" needs more conversation.

Acknowledgments: I am indebted to Robert L. Guyton for notice of Dreyfus's review article on Heideggerian AI, and to Cindy Mason for notice of the Conference.

References

Ball, J. 2012. *New Sprouts: A Haiku Anthology*. Lulu.

Berger, P., and Luckmann, T. 1966. *The Social Construction of Reality*. New York: Doubleday.

Berger, P. 1969. *A Rumor of Angels*. New York: Doubleday.

Damasio, A. 1995. *Descartes' Error: Emotion, Reason and the Human Brain*. New York: Avon.

Dreyfus, H. L. 1972. *What Computers Still Can't Do*. Cambridge: MIT Press.

Dreyfus, H. L. 2007. Why Heideggerian AI Failed and how Fixing it Would Require Making it More Heideggerian. *Philosophical Psychology* 20:247–268.

Dupré, L. 1993. *Passage to Modernity: An Essay in the Hermeneutics of Nature and Culture*. New Haven: Yale University Press.

Eliade, M. 1959. *Cosmos and History*. Princeton University Press.

[15] (Fingarette 1967), chapters 3 and 4 "To say or not to say," "To avow or not to avow."

Fingarette, H. 1967. *Self Deception*. Berkeley and Los Angeles: University of California Press.

Freeman, W. J. 1995. *Societies of Brains: A Study in the Neuroscience of Love and Hate*. Mahwah, NJ: Erlbaum.

Freeman, W. J. 2000. *How Brains Make Up Their Minds*. New York: Columbia University Press.

Heidegger, M. 1960. *Being and Time*. New York: Harper and Row.

Kierkegaard, S. 1941. *Fear and Trembling and The Sickness Unto Death*. Princeton University Press.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.

Mason, C. 2008. Human-Level AI Requires Compassionate Intelligence. AAAI Workshop on Meta-Cognition http://www-formal.stanford.edu/cmason/circulation-ws07cmason.pdf.

McCarthy, J., and Hayes, P. J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B., and Michie, D., eds., *Machine Intelligence 4*. Edinburgh University Press. 463–502.

Porter, A. P. 2011. *Living in Spin: Narrative as a Distributed Ontology of Human Action*. Bloomington: AuthorHouse. http://www.jedp.com/spin.

Reagan, C. E. 1968. Ricoeur's diagnostic relation. *International Philosophical Quarterly* 8:586–592.

Ricoeur, P. 1966. *Freedom and Nature: the Voluntary and the Involuntary*. Evanston: Northwestern University Press.

Wheeler, M. 2005. *Reconstructing the Cognitive World: The Next Step*. Cambridge: MIT Press.