THE NEW ATLANTIS
A Journal of Technology & Society

Robotic Souls

Charles T. Rubin

Today we see widespread interest in developing artificially intelligent robots as companions, caregivers, and sexual partners. Japan has become famous—but is hardly alone—for developing caregiver robots to deal with the oncoming deficit of its own citizens to look after an aging population. Just recently, *Scientific American* published an article titled "Grandma's Little Robot: Machines that can read and react to social cues may be more acceptable companions and caregivers." Surely it would be interesting to parse the significance of the caution implied by "may be."

Meanwhile, it seems an absolute truism among certain futurists and libertarians that robots are the next big thing in the sex trade. And indeed the creation of sex bots is underway. Some provocateurs have argued that these robots could help to resolve the sexual frustrations of lonely men, but the public has generally regarded these developments as concerning, laughable, or creepy. Nevertheless, the effort to create them is driven by powerful commercial motives.

At the same time, there seem always to be new impressive developments in the field of artificial intelligence—to name a few recent examples, self-driving cars, a program that plays Go at the highest level, and various high-quality medical diagnostic systems. These are admittedly not examples of what is sometimes called "strong AI," that is, AI that shows something like the full range of abilities of a human mind. But increasingly these narrow-application systems are developed through deep-learning techniques that are at least closer than previous methods to allowing AIs in effect to teach themselves—which suggests the possibility that far more widely ranging intellectual abilities could be developed.

In short, given the notoriously rapid rate of technological development, in the longer term it may well be that an effort to create an artificial human-like mind is not a fool's errand. Already it could be matched with a *virtual* "body" that under limited circumstances might be mistaken for

***Charles T. Rubin**, a* New Atlantis *contributing editor, is a professor of political science at Duquesne University, and the author of* Eclipse of Man: Human Extinction and the Meaning of Progress *(New Atlantis Books/Encounter, 2014). From 2017 to 2018 he was the James Madison Program Forbes Visiting Fellow at Princeton University. This essay has been adapted from a talk delivered at Princeton at the 2017 Robert J. Giuffra Conference of the James Madison Program in American Ideals and Institutions.*

human in an on-screen encounter. Such avatars will surely only become more convincing in the not-so-distant future.

Real embodiment, however, is farther off than is supposed by many of those working on it, as we can see in their tendency to fall prey to something like a Pygmalion syndrome when promoting their own, often not even remotely compelling, works. But there is little reason to doubt the ability of human ingenuity ultimately to triumph here as well. Creating a robot with a human-like mind in a human-like body would certainly be a great advance from the perspective of those who advocate a transhuman and posthuman future, a future where intelligence is no longer bound to the constraints of the organic body bequeathed to us by the random processes of evolution. But the drive for human-like robots does not, for the most part, depend on these aspirations.

Questions about the moral status of robots that (so we assume) would look and act in ways that make them hard to distinguish from human beings have been raised by popular accounts of robots from the beginning—the 1921 play *R.U.R.* that gave us the term "robot" was in large part concerned with the moral meaning of the exploitation of these artificial humanoids. Today the academy is beginning to catch up, under the rubric of asking whether robots will have rights.

Our answers to questions about the moral status of robots will depend in part on whether we can find any morally relevant grounds on which to distinguish robots from humans. Certain distinctions are made relatively often: between artificial versus natural intelligence, and between behavior that has the mere appearance of consciousness versus the actual possession of self-consciousness. But we might also do well to reintroduce what is today a somewhat less familiar category: soul. For thinking about souls would allow us to confront the challenges that human-like robots will present at least as well as, and probably better than, thinking about robots in terms of artificial intelligence or consciousness.

From Soul to Consciousness

Why, generally speaking, do people think about souls at all? Without meaning to slight the role of revelation, we might say that talk of the soul arises rather naturally out of various perennial human questions about perennial human experiences. How is it that we maintain a sense of identity despite physical changes over time? What accounts for our sense that we are wholes despite the manifest fact that we are collections of parts (psychic and physical) that, in truth, do not always work together? Most

fundamentally for our present purposes, we wonder how it is that we are different from cats, and cats different from stones.

We talk about soul because, first of all, we want some way to get at the fact that as animals, as embodied beings, we are, unlike stones, animate, and to that extent we in some way have *animas*. The Latin word for "soul" here supplies the placeholder for the ultimate, and not immediately obvious, source for why there is an obvious difference between living cats and rocks. For human beings, the situation seems yet more complicated. We are, to name a few distinctions we often point to, *unlike* other animate animals in our ability to make deliberate or intentional choices, to act creatively, to confound expectations, to be torn, to have immortal longings. So we have a soul that in some way—probably with respect to intellect—transcends the animal *anima* and allows us a certain kind of freedom. What this soul is could ultimately be to some extent mysterious, but something mysterious may yet exist. The soul could be like the cosmologists' "dark matter"—that is, we see and experience the *results* of soul all the time, even if a precise understanding of the thing itself remains elusive.

Just now, the soul is not an interesting concept for most philosophers, still less for scientists, and even many religious or "spiritual" people seem to have pretty much given up on it. But that does not mean that most of us have stopped noticing that cats are not stones and people are not cats. (Some are working very hard not to notice, it should be said.) It is just that today we try to explain the same kinds of experiences that led us to soul by talking instead about consciousness or self-consciousness.

We speak of consciousness instead of soul today not because the fundamental human experiences that formerly led to soul-talk have changed, but largely because, as the philosophers Raymond Martin and John Barresi have documented in their book *Naturalization of the Soul* (2000), modern philosophers wanted to give an account of human beings and human questions that was free of the mysteries of a soul presumed to be non-material. To some (for instance in John Locke's philosophy) the concept of consciousness was a kind of promissory note that in the future it would be possible to give a complete account of human things on purely materialistic and deterministic grounds. Human consciousness, like cats, stones, and everything else we observe in nature, ought to be explicable in terms of matter and motion. What we call human freedom, one of the sources of soul-talk, arguably then becomes a product of our ignorance of causes; someday we will come to see how illusory it is, and our immortal longings will be replaced by modern science's infinite task of determining

the causes of things. Consciousness promises to explain away many of the very things soul attempted to explain.

That day may be coming, but it has not yet arrived. (There is good reason to wonder whether it ever will: Patrick Lee and Robert P. George offer reasons for doubt in *Body–Self Dualism in Contemporary Ethics and Politics*.) People deeply schooled in the topic of consciousness argue vociferously about what it is and where it comes from. Supreme Court Justice Potter Stewart famously said of "hard-core pornography" that even if he could not define it, "I know it when I see it." Yet, as the lively debate over animal consciousness suggests, we are not all that sure we always know consciousness when we see it. The most telling indication of this impasse may be that now there are some, like Daniel Dennett, who in the face of these mysteries say that consciousness, like soul, is an illusion. We can only be quite conscious of the fact that we have little understanding of consciousness. To that extent, most of the mysteries that "the soul" was there to talk about—mysteries of the human way of being in the world—remain with us. We cannot yet cash the consciousness promissory note.

From Consciousness to "True" AI

Artificial intelligence steps into the breach created by our failure so far to understand consciousness. Most AI developers, however, have turned away from talking about consciousness at all. In doing so they follow the lead of Alan Turing, who separated the issue from intelligence in his famous 1950 essay "Computing Machinery and Intelligence," in which he wrote:

> I do not wish to give the impression that I think there is no mystery about consciousness. . . . But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper.

The question asked in the paper, following the behaviorist orientation then on the rise in psychology, was not whether there could be a conscious machine or even a thinking machine, but rather whether people could be *convinced* that a machine was thinking in a human-like way.

How do we know humans are thinking? Behaviorally speaking, because we can have a conversation with them. Hence the Turing Test, which Turing himself called the Imitation Game: A person is confronted with an interlocutor—in Turing's version, they exchange messages via a text-based chat—and must determine whether he is chatting with a computer or a fellow human being. (The object of the original game that

Turing derived his from was to tell a man from a woman.) If his interlocutor is in fact a computer but he thinks it is human, the computer has "artificial intelligence" by Turing's definition. If we are unable to clarify what consciousness is, then perhaps we are on firmer ground with defining artificial intelligence as that which is indistinguishable in daily life from human intelligence.

It could be said that AI is actually more or less the fulfillment of the materialist promise, which the switch to thinking about consciousness intended but could not achieve. We understand (more or less) the materialistic foundations upon which our computer-based AI is built, and it functions in a (more or less) deterministic way. AI promises to vindicate the Baconian idea that we know what we make. And, as things have turned out, AI that appears to think as we think by doing at least some of the things we do is all around us and quite impressive—autonomously driving cars and all-but-autonomously flying and landing airplanes, playing chess and computer games at the highest levels, winning *Jeopardy!* and developing recipes, taking prescription orders, providing customer service, correcting our spelling, finding restaurants and movie times. How many chat-based tech-support suppliers are people, and how many are chatbots? AI is already legion and looks to grow only more so. From this point of view we can understand author Yuval Noah Harari's assertion that the future will be molded by intelligence, with or without consciousness.

But many in the field would say that these successes have been won by abandoning Turing-style AI in some measure. Turing's own examples of human–computer interactions are premised on a machine that can talk like a human being who has had and retained a pretty good liberal arts education. Among AI developers, this effort has turned into a niche focus at best. Instead of trying to program computers to create foxes who know many things, most AI today attempts to create hedgehogs that know one great thing. The AI that flies your plane could never drive your car, nor could the AI that gives you directions drive your car. So far, the greatest AI successes have come by carefully defining the relevant domain of intelligence a given AI is designed to possess.

Yet there is a notorious problem created by this shift, which is nicely summarized in a 2016 interview with Yale ethicist Wendell Wallach:

> It has now become a bit more confusing what the term [“AI”] actually does and doesn't mean, largely because every time a goal is achieved, such as beating a human at chess, the bar gets raised. Somebody says,

> "Well, that wasn't really artificial intelligence in the way it beat the human at chess, in this case Garry Kasparov, because it didn't really play the way a human chess player would play."
>
> But even the folks in the more advanced fields of artificial intelligence feel today that we are just beginning to have true artificial intelligence, that a lot of what we have done so far is largely automating systems, largely programming them to follow through procedures that humans have thought about in advance.

In this understanding, an automated system–style AI lacks something that human intelligence has and "true artificial intelligence" would have. What might that be? One obvious difference, as noted above, is applicability over the broad range of functions and tasks an intelligent human can at least potentially perform. *Potentially* is the key word, however. We are not all equally good at doing everything that our fellows can do. There seem to be many types of intelligence, and many degrees of intelligence. What form and degree of human intelligence would we have to model to have "true" artificial intelligence?

Wallach says that automated systems follow routines that are the product of previous human thought. And yet much of the human knowledge we associate with intelligence arises only on the basis of what are, in effect, learned routines about which people are not necessarily very reflective or even very creative. If we adopt too stringent a definition of artificial intelligence, we may find ourselves excluding many forms of what we might otherwise call human intelligence. Would we say we have an artificially intelligent artist if it could explain itself as badly as the rhapsodist Ion does to Socrates (in the Platonic dialogue named for him), or would it have to do better?

Were it not for pervasive discussion of the "Singularity," the point at which artificial intelligence so far exceeds ours as to be incomprehensible to us, this high-toned view of creative and reflective "true artificial intelligence" that Wallach leads us to consider might suggest that an AI could educate and expand human intelligence. We would know we were being genuinely educated if this true AI could explain itself to us, could give an account of the fruits of its intelligence. Perhaps after all we should say that we had true AI if we could have a dialogue with it, if it could hold a conversation with a human being that would be like a conversation between two human beings. Contrary to appearances, then, the ghost of Turing could still haunt our search for the "true" artificially intelligent machines that go beyond automated systems.

Back to Consciousness and Soul

But the ghost of Turing is also the ghost of consciousness. If conversations with a machine suggested a self-understanding (or an obliviousness?) comparable to discussions with a real person, if it exhibited intentionality in its creativity (or cluelessness in its use of clichés?), if it understood its novel point of view *as* a point of view situated in relationship to *other* points of view (or was dogmatic and narrow-minded?), would we say it was *not* conscious just because we made it? A behavioral model yet more robust than that of Turing, who abstracted from bodily presence altogether, would have this much going for it: In practice, our preliminary judgment that we are dealing with a fellow conscious being is based on his or her embodied appearance, to which we grant the presumption of consciousness and so also communicativeness. So wouldn't the question of consciousness arise all the more if the machine could communicate with us in *all* the ways that human beings communicate—with tone of voice and body language, with all of the affect present when we encounter each other in the world, affect that depends upon our embodiment? All such characteristics that might convince us that we have "true AI" seem to force us to confront the question of consciousness again.

And if we reach consciousness, we are not so far from being back at soul. For it is only under the assumption of materialism and determinism that we substituted consciousness for soul in the first place, and that assumption did not get us as far as we hoped. We *could* conclude that, because a machine could appear to be very like a human being, a human being is nothing more than a "meat machine," as some of our transhumanists would have it. Or we could, in Leon Kass's terms in *The Hungry Soul* (1994), wonder about its soul—its "integrated vital powers," its "traffic with the world," the signs by which we see it to be creating a "lived space" or an "action space."

My intention is less to suggest that these as-yet-only-imagined human-like robots will have souls in some meaningful sense of the term than to point out why the question of their ensoulment is no less reasonable than the question of their consciousness or artificial intelligence. Indeed, thinking about soul is *more* reasonable to the extent that doing so allows us to address more directly the fundamental experiences that prompt the existential questions of our soulfulness to begin with—questions whose answers might even extend beyond our powers to reason about them. It is from this point of view about our machines that we would have the richest possible understanding of the human world of which they will be a part,

an understanding that extends beyond efficiency, convenience, choice and the other dogmas of our age, to question how exactly robots are going to fit into a well-lived human life.

This approach might start us along that path to wonder what it means that so many souls among us, and those not among the least powerful and influential, are longing to replace intimate human relations of care, love, and even pleasure with machine relationships. Unless we can take a question like that seriously, it seems likely we are setting ourselves up for a double failure in the coming world of robot caregivers and intimate partners. These relationships could turn out badly if in some manner these artificially intelligent machines end up disappointing their dependent human users for some eventually revealed lack of humanity. Or they will turn out badly if the machine never disappoints because it is *just* good enough, because our expectations for our relationships have been reduced and narrowed *just enough* that the very satisfaction gained from the machine relationship forecloses any desire for the complex possibilities of human relationships.