Turing, Searle, and the Wizard of Oz: Life and Custom Among the Automata or How Ought We to Assess the Attribution of Capacities of Living Systems to Technological Artefacts?

S. D. Noam Cook
San José State University

Abstract
Since the middle of the 20th century there has been a significant debate about the attribution of capacities of living systems, particularly humans, to technological artefacts, especially computers —from Turing's opening gambit, to subsequent considerations of artificial intelligence, to recent claims about artificial life. Some now argue that the capacities of future technologies will ultimately make it impossible to draw any meaningful distinctions between humans and machines. Such issues center on what sense, if any, it makes to claim that gadgets can actually think, feel, act, live, etc. I outline this debate and offer a critique of its persistent polarization. I characterize two of the debate's major camps (associated roughly with Turing and Searle); argue that the debate's structure (including key assumptions inherent to each camp) precludes resolution; and, contend that some central clashes within the debate actually stem from an inadequately drawn distinction between claims about the capacities of artifacts and claims about the proper criteria for assessing such attributions. I offer a different perspective in which I: challenge some central elements of the debate that contribute to its perennially irresolvable state; hold that the debate needs to be placed more squarely in sync with how we in fact treat the attribution of such capacities to humans themselves; and, offer (unlike the other two camps) a foothold for making moral assessments of such proposed technologies.

Keywords: Turing Test, Chinese Room, Artificial Intelligence Debate, AI Debate

1. Introduction

Whatever else the technological may be, the imposition of human design on the material stuff of nature seems inescapably at its core. Technology is about the mingling of the made with the given. And although this mingling is common in human experience, and may be inextricably tied to the human condition, we do not systematically fail to distinguish between the two. It is important that we do not, since the given and the made have different requirements for use and sustenance.[1] Indeed, effective and responsible use of technologies requires that we honor this difference, even though (or especially because) the two are embodied in a single artifact. Without viable criteria for distinguishing between the given and the made, we are in serious trouble, both instrumentally and morally.

These points are nowhere more relevant than in the debate about what computers can and can't do —or, more properly, about what capacities computational systems may be understood to have. Central to this debate is the issue of whether computers might ever possess capacities that traditionally we have understood to be possessed only by living systems, particularly by humans. These include the capacities of thought, intelligence, consciousness, feelings, and life. Such topics raise important questions about how we ought to understand the relationship between the

given and the made, and about the criteria by which we ought to be able to distinguish between them—especially the need to distinguish between our artifacts and ourselves.

This debate has raged on since the middle of the 20th century in many forms and on many themes of great complexity and sophistication, both technical and philosophical.[2] Here I make some general observations about the debate itself, and venture a few observations focused on the importance of maintaining our ability to draw distinctions of the sort I've just mentioned. I will begin by looking at two major camps in this Great Debate, one in the tradition of Turing, the other in the tradition of Searle.[3]

2. Turing

The noted mathematician Alan Turing published a now-famous essay in 1950 in the journal *Mind* titled "Computing Machinery and Intelligence" (Turing 1950). In the article, Turing considers how we might determine whether or not a computer could think, and proposes a test for making this determination. The Turing Test, as it has come to be called, has been widely used, cited, debated and misunderstood ever since. The Test is generally taken to go something like this: put a human in one room and a computer in another, then have another person, whose only contact with the rooms is via two terminals, ask questions of both in a given field; if the person cannot reliably tell the human and the computer apart based on their answers alone, then we must conclude that whatever capacity is being tested by the questions is possessed by both. In other words, if the computer can engage in a discussion in a way that is indistinguishable from a parallel discussion with a person, then the capacities demonstrated by the person must be possessed by the computer as well. From the time Turing's article appeared to the present day, variations on this test have been given in support of various claims about the capacities of computers. And a growing number of people have associated themselves with this tradition as the proper way to think about computers when assessing their capacities. (Whether Turing's article, itself, actually supports this position is an issue I will return to later.)

No sooner had Turing's article appeared than it began to draw a series of critiques. A dominant theme that emerged among them was that there *must* be something more to thinking than just a mechanical procedure tucked away behind the ability to answer questions—*surely* thinking must be something more intimately tied to who and what we are as living, biological beings.

3. Searle

The philosopher John Searle offered a rebuttal to the Turing camp, broadly speaking, in 1980 in the journal *Behavioral and Brain Sciences*, which has drawn its own set of followers and critics. (Searle, 1980). The Turing camp, in Searle's view, looks at the wrong thing: machines may *appear* to have human capacities like thought or consciousness, but this is insufficient proof that they actually do. In making his case, Searle drew an analogy now commonly called "The Chinese Room." Here we are asked to imagine a person in a room who is given a book of instructions and a stack of cards with what are to him unintelligible squiggles on them. Every once in a while a new card with a squiggle on it comes through a slot in the door. The person matches the squiggle on the new card to a picture of it in the book. Next to the picture he finds a picture of another squiggle. He then finds a card from the stack with the second squiggle on it and slips it out the slot in the door. This goes on for some time. Later he learns that the squiggles on the cards are Chinese characters, and the person who had been slipping cards through the slot all afternoon thought she had been having a conversation with him in Chinese. What actually was going on, Searle argued, was that the instruction book enabled the man to match up questions and answers

in Chinese without knowing a word of the language. The woman outside the room, accordingly, would have been mistaken to think the man inside knew Chinese.

For this camp, the Chinese room is the equivalent, for all relevant purposes, of the computer. Computers take in symbols, deal with them according to instructions in their programs, and generate appropriate output. But they do not know what the symbols mean. Computers do not have the capacities the Turing test makes them appear to have because they are, like the man in the Chinese Room, ignorant of any content that their overt behavior might suggest. Computers, this camp argues, may have the syntax, but they lack the semantics. And without content or understanding, there is no thought or consciousness, only mechanical manipulation of symbols. It makes no more sense to say that a computer can think than it does to say the man in the room can speak Chinese. And there is a further and important reason for this. The content of our thoughts, the meaning inside human thinking, this camp maintains, is intimately tied to the fact that we have the sort of brain that we do. Put another way, our subjective experiences of thoughts, sensations, emotions, and the like (what are traditionally called "qualia") are produced by activities of our brains: we experience these things because we have the sort of brain that can carry out the kinds of activities that cause them. Without the sort of thing that our brain happens to be, there is no thought or consciousness. A robot might be made to *behave* as though it were conscious, but it would not *possess* consciousness or thought, it would be nothing more than a mechanical zombie.

The Turing camp has made many responses to the Searle camp, which in turn has responded back. And this noisy and volatile bit of the Great Debate goes on to this very day.

4. The Wizard of Oz

L. Frank Baum's story *The Wonderful Wizard of Oz* was first published in 1900, and has remained an icon of American literature. MGM's 1939 movie "The Wizard of Oz," directed by Victor Fleming and staring Judy Garland, was based on Baum's book and is a classic of American film.

The film begins with Dorothy living in Kansas, which is flat, boring and filmed in black and white. A cyclone soon carries Dorothy, her house and her little dog Toto off to the land of Oz, which is full of fanciful characters, beautiful settings and is filmed in Technicolor. Immediately upon her arrival, Dorothy declares her interest in going back to Kansas. (I never understood this part.) She is told to visit the Wizard of Oz, who will surely be able to send her home. On her way to the Wizard she meets up with three characters, each of whom decides to go with her to the Wizard and to make a request of his own. There is a Scarecrow, who longs to have a brain so he could think great thoughts; a Tin Woodsman, who started out as a human but through a series of accidents with his axe ended up having all his original body parts replaced with tin ones, and who now wants a heart, so he might experience feeling; and, a Cowardly Lion, who, naturally, wants courage so he can become king of the forest. These things alone would make the Oz story a good candidate for inclusion in the Great Debate, at least as an allegory. But, for the moment, I have something else in mind.

The Wizard agrees to grant their requests, if the four bring him the broomstick of the Wicked Witch of the West. With some difficultly they get the broomstick, bring it back to the Wizard's throne room and ask to him to keep his part of the bargain. The special effects at his point are fantastic, especially for 1939: the image of the Wizard floats in a cloud of smoke with shooting flames and flashing lights while his thundering voice barks at our foursome who are trembling before him. Meanwhile, Dorothy's dog Toto scampers to the side of the room and brushes back a

curtain. This reveals an unimposing man madly working a bank of levers and wheels while speaking into a microphone. Dorothy and her companions see the man and realize that he is all there really is to the Wizard. At the same moment, the man, upon seeing that he has been revealed, pulls the curtain closed again and utters into the microphone, as the voice of Oz, one of my favorite lines in all of film history, "Pay no attention to that man behind the curtain."

I would like to believe there is a lesson in this that may have some applicability to the Great Debate. In particular, I would like to revisit the Turing Test and the Chinese Room to see if there is anything in them that we are being asked to "pay no attention to."

It would seem that those in the Turing camp are behaviorists. Their central argument points to behavior as the proper criterion by which the capacities of machines are to be established. If, under certain conditions, a machine can behave in a way that we normally associate with a thinking person, then the machine must be taken to be capable of thought as well. But this asks us to pay no attention to anything except behavior. The computer, as a machine, a gadget, a program, a bunch of printed circuits, etc. is, in effect, kept behind a curtain: we are asked to treat the fact of what the computer *is* as something that it "merely happens to *be*," and consequently, as something not relevant to the questions at hand.

Yet, there is also an interesting sense in which it could be argued that those in the Turing camp are more materialists than behaviorists. This would rests on their long-standing insistence that any behavior in question must be physically instantiated. A description of behavior encoded in software would not by itself qualify for them as thinking: the encoded behavior has to be materially manifested. However, to the extent that this is a materialist claim, it is only incidentally so because the Turing camp does not require that the material have any particular properties, except that it is capable of acting as a substrate for the behavior in question. Indeed, Searle once queried whether the AI people would accept the possibility of a thinking machine made out of beer cans. Although this may have been intended as a rhetorical gambit, the response of many AI adherents has been, basically, "why not?" Thus, the Turing camp asks us to pay no attention to the nature of the material, and offers behavior as their favored criterion.

If those in the Turning camp are behaviorists, members of the Chinese Room camp appear to be essentialists. For them the behavior of the machine is something that it "merely happens to *do*." The important factor for them is the nature of the machine itself—that is, what the machine *is*. No machine so far proposed can be a conscious and thinking entity no matter what it does, they insist, because all of them are simply the wrong kind of thing. (There is one theoretical exception to this, to which I will return later.) Here, basically, the curtain is drawn over behavior.

At this point it can be noted that while at one level the Great Debate is about the capacities of computers, at another it is about the correct way for determining this: what passes for an argument about what computers can and can't do is often at root a clash over which criteria are deemed appropriate for determining this (e. g., behavior versus essence). Yet, this distinction itself is frequently obscured. Failure to make it clear, I believe, has often been an unidentified source of confusion in the Debate. Further, it seems to me that the criteria offered by these two camps are themselves ultimately incommensurable: there is no guarantee that any amount of additional detail about behavior will ever offset all possible objections based on essence, and vice versa. So, to the extent that the Great Debate remains polarized in this way, it could go on forever —as the last several decades might portend.

The origins of this incommensurability, by the way, may lie in part in the dominant traditions of the two camps and the "habits of thought" common to them. The dominant tradition of the Turing camp is science. There is a significant focus in the sciences (and in the training of scientists) on modeling behavior. Gravity, students of science learn, is defined as that which causes things to behave in various ways, which are wonderfully modeled in the calculus, predicted by Newton's equations, etc. Scientists (including Newton) are not particularly concerned with what gravity actually *is*, but rather with how physical objects behave in its presence. Meanwhile, for the Chinese Room adherents the dominant tradition is philosophy, where, it should be confessed, it is all too easy to see a predilection for focusing on essence.

So, let's pull the curtain back a bit further. As Alan Turing describes it in his seminal essay, the Turing test has its origins in a Victorian parlor game. In the game, an interrogator, via written questions and answers, tries to guess the sex of two hidden individuals, a man and a woman. The aim of the game, however, is to try to have the interrogator make an incorrect guess. Accordingly, the woman, say, must answer all questions honestly, while the man answers them with the intention of making the interrogator guess that he is a woman. Substitute a computer for the man, as Turing suggests, and you have the Turing Test.

But what does the original parlor game actually tell us? The game may trick us into guessing that a man is a woman, but in doing so it proves nothing about the man and the woman themselves. At the very least, the parlor game gives us no grounds whatsoever for claiming that the man is *in fact* a woman. Likewise (to keep the logic parallel), the Turing Test, as defined by Turing himself, would seem to give us no grounds for concluding that a computer can *in fact* think. That we can be tricked into guessing that it thinks does not support any claims that it actually does. At best, the original parlor game demonstrates that within the context of the game the man can successfully imitate a woman (Turing, in fact, called it "the imitation game.") Put another way, the game can enable the man to fool us. (That the game can lead us to draw false conclusions is, alone, reason to be suspicious of it as a model for a reliable test.) Significantly, whatever this may tell us about the man, we should not ignore what it tells us about the game: it is the set-up and restrictions of the game that enable the ruse. Thus, the ruse does not prove anything about the man outside of the context of the game (with the curtain pulled back, so to speak). To keep the logic parallel once again, it would seem that the Turing Test gives us no basis for concluding anything about machines outside of the context of the Test itself: demonstrating that within the Turing Test a computer can imitate thought does not give any grounds for concluding that it can actually think in general. Using the Test to do so, it seems to me, is not a test but a trick. Thus, I call this camp the "Turing Tricksters."

To be fair in handing out names, since the adherents of the Chinese Room tradition so unfailingly stick together around their cause, I shall call this camp the "Chinese Roommates."

At this point, a useful lesson from the Oz story would be always to ask what criteria are being posited as appropriate ways by which machines can be understood to have traditionally human capacities. That is, which criteria are we, explicitly or implicitly, being asked to use, and which are we being asked to "pay no attention to"?

Given the commitment I am proposing here, always, like Toto, to seek out curtains to pull back, I suggest those inclined to this third camp be called the "Friends of Toto." Below I consider some key themes of the Great Debate and offer critical remarks concerning them from the perspective of the Friends of Toto.

5. Distinguishing Between Humans and Technological Artifacts

It has been asserted in recent years with increasing fervency that advances in science and technology, from computation to bioengineering, are making it ever more difficult to distinguish between humans and machines. Indeed, some have argued that eventually we will no longer be able to do so. I must point out, however, that if technology ever delivers us to the point where we cannot distinguish between humans and machines, by definition, we will never know it. This is an inescapable consequence of what being "unable to distinguish" means. Yet, this does not seem to be what these claimants have in mind. Typically, they hold that we will not be able to make this crucial distinction under some "given conditions." Again, there's the rub: those conditions invariably entail putting *something* behind a curtain. The "given conditions" imply (explicitly or implicitly) that certain things are proper criteria for making such distinctions and others are not. Yet rarely, if ever, is any substantive argument given for why this is so. Thus, there is nothing new here: we are back to the same situation we were in with the Tricksters and the Roommates.

6. Models, Simulacra, and the Real Thing

The Great Debate often entails the use of models. Here a central question is whether something about a computer actually constitutes the machine's possession of thought, intelligence, emotion, etc. or is nothing more than a model of it.

That we can make things that model human capacities seems clear enough. And such models need not be sophisticated at all. Indeed, there have been machines that could model such things as dealing cards and writing with a pen that date back now a century or two. But just as a map is not the territory, positing the likes of a computational artifact as a *model* of intelligence, consciousness, emotion, or the like does not in itself establish that it actually *possesses* that capacity.

So how might we understand models in a way that is applicable to the present topic? It is my contention that in a very general sense, every single thing can be offered as a model of any other thing. Almost all of them, however, would be very bad models. The contours of the south rim of the Grand Canyon could serve as a model for weather patterns over Bora-Bora, just not a very useful one. Whether a model is a good one or not depends on one's criteria for what I call the "appropriateness of fit" between the model and the thing modeled, with "appropriateness" being relative to the task at hand. Thus, the focus shifts once again to *criteria*. Determining whether a 19th century card dealing machine or IBM's chess program Deep Blue is a better model of human thought requires proposing criteria by which this can be gauged. And in any case, even if we were to settle on criteria for X being an utterly splendid model of Y (e.g., the characteristics of some machine being an appropriately fitting model of thought) we are still left only with a model, even if a very good one. That is, agreeing that we have a very good *model* of a human capacity is in itself insufficient grounds for concluding that we have an *instance* of it. This would seem to remain always true, unless we can show that a given model of a capacity does something more than simply modeling it.[4]

There is one case that some feel makes a critical shift in this respect—specifically, that of a particular kind of model called a computational "simulacrum".[5] The argument goes something like this: given certain criteria (again, not always argued for), you can build a model that simulates a human capacity (such as thought) with ever greater closeness, such that at some point the two will become so close that any relevant differences between the model and the real thing will disappear. That is, the model will become an instance of the thing modeled: the computer

will *in fact* think, feel, etc. On a closer look, however, this is little more than an assertion. The great excitement with which it is often made, and the amazing technologies with which it is associated tend to distract us from this fact (in a way suggestive of special effects or a man behind the curtain). Ultimately, the simulacrum position lacks in substance exactly what on the surface it proposes to add: it does not show (or often even *attempts* to show) just when or how exactly we get to a point where the model and the real thing become equivalent; it does not shine sufficient light on why the simulacrum does anything more than simulate. The spirit of this position seems to be "*surely* at some point the right kind of model *must* become identical to the real thing." Yet no amount of hand waving alone will make this so. The implicit appeal, I think, is to the notion that increasing quantity must ultimately of necessity beget a difference in quality. But this is not obvious in general, and it certainly is not clear in this case.

In any event, any argument for simulacra would have two very thorny issues to address: there is nothing logically inconsistent in the idea of a model that more and more closely simulates a human capacity yet remains nothing more than a model; nor is there any necessary logical limit to how close the model can get to the real thing *without* becoming an instance of it (with one exception that I will come to shortly). Thus, for the simulacrumists (if I may call them that) to make their case, they would at minimum need to show that a given model is a true simulacrum and not merely what we might call a super high-fidelity simulation. Whether we are trying to climb up to the ultimate plateau or swim down to the last turtle, this is the sticking point for the proponents of computational simulacra: by what criteria do we distinguish between a simulacrum that *in fact* thinks, feels, etc. and something that may be an extremely good simulation of these human capacities, but merely a model nonetheless? We await the answer.[6]

In passing, it is perhaps worth noting that one possible source of the inclination to see finer and finer simulations as ultimately becoming the real thing may stem, in part, from the central role played by mathematics in the sciences, including computer science. (I mean this in a way akin to how the "habits of thought" in science and philosophy noted above may incline scientists to favor behavior while philosophers favor essence.) Mathematically, we have no trouble at all stacking up an infinite number of infinitesimal increments to reach a limit in a finite bit of time and/or space. Importantly, this can yield fantastically powerful tools for modeling the behavior of physical systems. What happens with a ping-pong ball, for example, between the time it is dropped and when it comes to rest is wonderfully modeled in this way. But it is quite another thing to assert that physical systems so modeled actually have the capacity to do an infinite number of things in a finite amount of time. In practical terms, there is absolutely nothing that we know about physical systems that supports the notion that a ping pong ball can bounce an infinite number of times under *any* circumstances (even if given an eternity to do so). Although mathematically we can tell a wonderful story about how a bouncing ping-pong ball transforms into a stationary one, like reasoning does not enable to us to prove that an increasingly hi-fidelity simulation will transform into an instance of what is it simulating.

The Chinese Roommates, meanwhile, might address the question of distinguishing between a high-fidelity simulation and the real thing in a way that could reveal a telling twist implicit in the Great Debate. As I alluded to above, the Roommates have not categorically ruled out the possibility of computers thinking or being conscious. This could happen, they say, if we could build a machine that has the same kind of physical properties that enables our brains to find meaning in the world and to act on that meaning. This follows, in fact, from a central assertion of their position: the reason we have subjective experiences such as thoughts, emotions and sensations, is because our brains have physical characteristics that operate in ways that cause

them. Thus, if we could discover what these characteristics are and could build a machine that has them, then (the Roommates say) the machine would think and feel as we do.[7]

Given this stance, we might ask the Roommates how we ought to distinguish between such a machine and a mere machine model of the physical characteristics they identify as the necessary underlying physical conditions of consciousness and thought. For the Roommates, however, the question is most likely moot. Indeed, in a key sense, they would have to hold that such a model would be impossible. If one claims, as they do, that certain physical properties of the brain are sufficient to cause thought, feeling, etc., then it follows that if one were to succeed in building a duplicate of those causes, one presumably would get the same effects—that is, ultimately the model must necessarily cease being a model and become an instance of the real thing. The Roommates (like the simulacrumists) may be unable to say exactly at what point (or across what range) this happens. However, they do (*un*like the simulacrumists) have something in addition to ever-closer simulation that says where consciousness and thought come from—namely, they come from particular physical properties (of brains or artifacts built like brains).

On what, then, the Friends of Toto would ask, do the Roommates base their claim about these proposed causal properties of the brain? The answer (and here begins the telling twist) seems to be nothing more and nothing less than that the Roommates assume it is so. And just as they assume it about human brains, so do they assert it about potential artifacts with the same relevant physical properties.[8] Yet, it has never been proven either conceptually or empirically that particular physical properties of the brain are all we need in order to have the subjective experience of thought, feeling, etc. Nor has this been shown even to be provable (or un-provable). If we assume it, certain things follow—even important things. But what follows cannot make the assumption any less of an assumption. Yet, on this assumption the whole of the Chinese Room argument seems to rest. Indeed, once you have made the assumption, the Roommates' argument follows quite elegantly.

Moreover, when it comes to making such assumptions, the Roommates are not alone. The Tricksters, too, make a big assumption from which the core of their position follows—namely, if the *right kind* of behavior is instantiated in *any* material thing (brain or machine) that can serve as a substrate for that behavior, then consciousness and thought will ensue.[9] This assumption is also unproven, and may be un-provable. Significantly, both assumptions lie at the heart of the Great Debate. Without them it would not arise—or it would be something else entirely.

Curiously, however, the Debate has been carried out substantially at the level of clashes between implications of each side's assumptions, while the assumptions themselves have often been left untouched or brushed over. This oddly ignores the likelihood that many of the implications clash precisely because they rest on two different and incompatible assumptions. Meanwhile, the tendency to debate at the level of implications seems in proportion to the conviction with which the assumptions on both sides are held—often to the point of treating them as self-evident fact. Such conviction seems to confer to the implications a sense of solidity that they in truth do not have.

The assumptions do indeed broadly support two sets of provocative implications, some worked out with remarkable care, some yielding challenging insights. Nonetheless, they remain logical implications of assumptions. And treating implications, no matter how carefully or insightfully wrought, as conclusions amounts to little more than intellectual slight-of-hand. Losing sight of this, I believe, has played a key role in why the seemingly logical arguments of one camp repeatedly fail to upset the convictions of the other, why each camp genuinely finds the

arguments and passion of the other bewildering, and why the Great Debate seems never any closer to resolution (especially to innocent bystanders).

As an aside, I would note that this standoff is not the only reasonable course open to the two camps. There is at least one other strategy that might yield some fruitful middle ground, even if it cannot address the attachment the Roommates and Tricksters have to their respective assumptions. Rather than endlessly drawing on the implications of one assumption to dispute the implications of the other, each camp could argue that its assumption is a logical requirement of the other's assumption. Thus, the Tricksters' argument would be that any artifact possessing the sort of intentional consciousness on which the Roommates insist must necessarily engage in the kind of behavior the Tricksters assume to be key. The Roommates' argument, meanwhile, would be that any material artifact that can act as a substrate for the sort of behavior (real, not just modeled) on which the Tricksters insist must necessarily have the causal properties the Roommates assume to be key. Yet, it remains to be seen whether either case would enable us to determine if or when or how a model might become an instance of the thing it is modeling (I will return to this later).

In the film version of the Wizard of Oz, after Toto exposes Oz for the ordinary human he is, Dorothy and her companions, disappointed and angry, confront him. At the height of their tirade, Dorothy snaps at Oz, "You're a very bad man." To which Oz gently replies, "Oh no, my dear. I'm a very good man. I'm just a very bad Wizard."

Even when we are trying to understand what something *might* be, there would still seem to be some wisdom in not losing sight of what it really is.

7. Androids and Cyborgs and Replicants, Oh my!

Whether they are drawn from leading edge research or from science fiction, human-like androids, cyborgs, and replicants raise another issue relevant to understanding the nature of the Great Debate. Such examples often appear in challenges like, "Suppose you *could* make a machine that is conscious?" or "What if a computer model *really were* totally isomorphic to a thinking brain?" or "What about the replicants in the film *Blade Runner*?" In general, I am a great fan of such questions. And *Blade Runner* is one of my favorite films. And I agree with Hannah Arendt that the philosophical value of science fiction is quite under appreciated. However, since I'm speaking here for the Friends of Toto, I have to insist that such examples put far too much behind a curtain, although in a different way from the Tricksters and the Roommates. Basically, these examples beg the question. They amount to saying "If you *really* were unable to distinguish between human and machine intelligence, wouldn't there be no difference?" or "If we *really* could completely simulate a human brain, wouldn't the simulation think?" or "If you *really* found it impossible to spot an android hiding among the humans, wouldn't it be just like us?" An essential response to rhetorical assertions like these is that what we *really* can't do is use hypothetical cases of this sort as grounds for concluding anything, one way or the other. For the purpose of making headway in the Great Debate, these examples, as useful as they may be for other matters, are simply of no help.

This problem also arises in the case of projections about future technologies based on current trends. It is claimed these days, for example, that intelligent, conscious and even "spiritual" machines will be made possible by future technological advances. However, we do not in fact know what advances will happen and which will not. So, just as the unknowable future has moved some to argue that such machines will one day exist, it must also require them to admit

that they might not. Like appeals to androids, cyborgs and replicants, speculation about the future of technology ultimately gives us no basis for concluding today what may or may not possible tomorrow.

8. What Computers Really Do

I have to note at this point that, from the perspective taken here, the bulk of this decades-old debate about what computers can and can't do has *not* been about what these technologies themselves *in fact* do. IBM's Deep Blue is a prime example. The debate about whether Deep Blue can play or understand or think about chess is not, in significant ways, about what the machine *does*, but rather about how we ought to interpret what it does, and about the conclusions that those interpretations seemingly support concerning the capacities of machines.

It is worth looking at what Deep Blue actually does at its most uncontroversial levels. At one level, Deep Blue takes in information that we can see as being symbols representing the deployment of chess pieces; it runs this information through some remarkably sophisticated procedures, which were designed as representations of the rules of chess and patterns of possible chess moves; it puts these results through other procedures, which were designed to apply an optimisation scheme to them; it then produces output that, again, we can understand as being about the reconfiguration of chess pieces. At an even more basic level, of course, all Deep Blue does is to pump huge volumes of electrons around at breathtaking speeds in devastatingly complex patterns. Neither of these descriptions is particularly problematic. Indeed, both the Tricksters and the Roommates could easily agree to them. In a technical sense, meanwhile, there is nothing in either of these descriptions of things that Deep Blue *really does* that requires us to talk about Deep Blue either "understanding" or "thinking" about chess. In fact, it is logically possible that we could make sense out a good deal of the above, maybe even all of it, without reference to chess at all; we might even be able to describe it in terms of something else entirely. This suggests that much of the debate about Deep Blue and other machines doing things like "thinking" is not actually about what they really do, but about how we ought to understand or interpret what they really do. (This may not apply to the whole of the Great Debate, but it applies to far more than is generally recognized.) It is at this point that the bulk of the controversy about such machines as Deep Blue begins.

The Deep Blue case, itself, might look like a useful one for the Chinese Roommates (at least at first). The claim would be that Deep Blue, like the guy in the Chinese Room, does what it does without attaching meaning to anything at all. The Roommates could be genuinely delighted for us to describe Deep Blue in terms of something other than chess, since this would demonstrate that a single syntax could be the substrate of two distinct bits of semantics. For the Tricksters, meanwhile, this claim could appear unduly stark. They might argue that any entity that can exhibit such complex behavior that maps so well onto chess must be understood to possess the capacity in question. Surely, they might insist, the Roommates are reducing semantics to a shell game. But this does not get the Tricksters completely off the hook since no one has yet to disprove the logical possibility of interpreting a given formal system in more than one way—for all we know, Deep Blue may one day be shown to predict those elusive weather patterns over Bora-Bora.

The Tricksters, ever undaunted, might counter attack. In this spirit, some have in fact argued that the Chinese Roommates have misunderstood the very contrivance they have invented. The poor fellow in the Chinese Room, the Tricksters admit, indeed knows not a word of Chinese. But *he* is not the parallel to the computer, they contend. Rather, the room as a whole, the system of rules

and procedures and the people carrying them out, the Tricksters claim, does in fact know Chinese. The Chinese Room itself, they say, is equivalent to a computer. But, Toto's vigilant Friends would insist, if this argument were right, then it would be fair game to apply the same reasoning to Deep Blue. This could have us conclude that it is not the computer, but the whole Deep Blue set-up, crew and all, that knows chess. However, this seems to be exactly what hard-core Deep Blue fans do not want us to think. "Pay no attention," they seem to be saying, "to the team of IBM programmers behind the curtain." Thus, the Great Debate goes on.

9. "If I only had a brain, a heard, the nerve..."

So, how might the Friends of Toto approach the issue of assessing the capacities of the machines in question?

It is perhaps an easy thing to see Dorothy's three companions as representing different parts of ourselves. It is not just a brain or a heart or courage that makes us human; it is all three, and a whole lot more. Hiding any one of them behind a curtain masks the whole of who and what we are (and long to be). In making sense of our lives and each other, we attend to behavior *and* essence, to thoughts *and* feelings, and so on—both in ourselves and in others. We are all the time, in ways both subtle and profound, trying to figure out what one another is about.

In this spirit, and on behalf of the Friends of Toto, I propose that if we wish to determine whether or not a technological artifact possesses capacities traditionally understood to be capacities of humans, we ought to use the same criteria by which we make such judgments with respect to humans themselves. There is, however, some difficulty in doing this. We do not know fully what these criteria are. Nor, in at least one important sense, are we able to nail the matter down definitively: it is a well-worn tradition in philosophy that I cannot prove beyond question that the person I'm having lunch with is intelligent, conscious, thinking, feeling, and the like.

Turing, himself, refers to this; indeed, it becomes a pivotal point of his famous essay. Since we cannot prove that other people think, to avoid becoming solipsistic, Turing says, we have developed the "polite convention that everyone thinks." He then offers his version of the imitation game as a parallel in the case of machines to our "polite convention" in the case of people. In doing so, however, all Turing claims is that if there comes a time when we are unable to distinguish between people and certain machines, at least under certain conditions, at least most of the time, we will tend to attribute thought to them. He offers nothing else, either conceptually or empirically, as criteria for dealing with the question of machine thought. Nor is there anything in his essay that would help us distinguish a thinking machine from a machine model of thinking. Nor is there anything to suggest, as noted above, why his game should have us conclude that a machine thinks, when it does not move anyone to conclude that a man who can trick us, through the contrived polite convention of the game, into believing he is a woman is *in fact* a woman. If one looks carefully at the arguments in Turing's essay, the fact that the Turing Test has become accepted in so many quarters as a valid gauge of the capacities of computers is rather baffling. At the very least, its acceptance as such cannot be explained by the content of the essay itself.

I suspect that adherence to the common understanding (or misunderstanding) of the Truing Test rests for many on a commitment to what I would call "behavioral symbolic reductionism." The underlying idea here is that a capacity like thought or feeling can be duplicated by constructing a symbolic (i.e., computational) representation of it, which can cause, in an appropriate material substrate (e.g., a computer, robot, etc.), behavior associated with the manifestation of that

capacity in a living system (e.g., a person). As noted above, the only criterion this approach requires for material to be "appropriate" is that it is able to serve as a substrate for the behavior. Based on this, the assertion is made (at least implicitly) that reducing such a capacity to a materially enacted symbolic representation of behavior associated with it (whether overt conversation, the functional patterns of neural nets, etc.) results in a true instance of that capacity —e. g., a machine that really thinks and feels. Yet, at root this assertion is nothing more than a metaphysical commitment, and as such it may be more indicative of tribal loyalty than of good science or good philosophy. There is much to be said about metaphysical commitments and tribal loyalty, both positive and negative. But one thing they do not enable us to do is resolve questions about the capacities of machines in a scientifically or philosophically sound way. (The passion with which others reject behavioral symbolic reductionism, meanwhile, may well reflect tribal loyalties of their own.)

Nonetheless, in his pivotal use of the idea of "polite convention" and his proposal for a parallel in the case of machines, I think Turing was on to something that has generally been under-appreciated. I don't think Turing goes far enough with it, however. So, I would like to propose something that might move us further along. I agree with Turning (and almost everyone else) that in the final analysis we cannot prove that other people are conscious or thinking or feeling happy, and the like. Yet, in an equally important sense, we make such attributions of our fellow humans all the time. In order to interact with another person in an effective and responsible way, there are many things I must assume about him or her. To converse with someone, to be a friend, to trust or be trusted, and to do such things in ways that allow my expectations of our interactions to be satisfactorily fulfilled and mutually acted upon, I must assume that the other is alive, a member of my species, intelligent, autonomous, capable of feelings, and a lot of other things as well. Indeed, I must also even assume that another's feelings, intentions, aims and sensitivities are for him or her rather like what mine are for me.

Again, I cannot prove these things in any definitive conceptual or empirical way, but neither can I act effectively and responsibly without assuming them (at least implicitly). I also judge the content of my assumptions by attending to what others say and do, to how they deal, in many ways, with me and with the world—which, in turn, I assume reflects like assumptions on their part, and so on. Effective and responsible interaction, indeed the constitution of all forms of human interaction in any form that we are likely to find acceptable, would be impossible without our making such assumptions. If this is how *in fact* we actually assess the capacities of other people, if these are indeed the criteria for determining, as best we can, that others possess what we inescapably find in ourselves, then I argue that we ought to apply the same criteria to artifacts that are put before us as possible possessors of such capacities. That is, we should apply to machines not this criterion or that, but the same range of criteria that we apply to humans, *in toto* (so to speak).

This formula I modestly call "The Cook Test," and I make a mild-mannered plea for its considered inclusion in the Great Debate. Accordingly, should we ever find ourselves dealing with life and custom among the automata, our understanding of their capacities would be revealed in the assumptions that we found we *must make* in order to interact with them effectively and responsibly.

10. "There's no place like home."

In a sense, I have been arguing that we need to put some key questions of the Great Debate into the context of ordinary life—into the place where we weave the various practices that constitute

the world in which we live; where we are most at home. We often forget (though there are philosophers and scientists who occasionally remind us not to) that there is no place like ordinary life, no better place anyway, to test our philosophical and scientific claims.

In focusing on what we must assume in ordinary life in order to interact with other people "effectively and responsibly," I have deliberately paired the instrumental and the axiological. Morally speaking, I would not be a *responsible* person if my interactions with others did not reflect my assumption that they have thoughts, feelings, and aims that are rather like my own. Nor could I be an *effective* friend, neighbor, colleague, etc., if I did not interact with others *responsibly* in this way (at the very least, I could not expect reliable responses from others if I were to treat them merely as gadgets). In our dealings with one another, acting effectively requires that we also act responsibly, and vice versa—neither functions reliably without the other; neither one should be put behind a curtain. In assuming that other humans are conscious, that they think, feel, and so on, we are not only saying that they have those capacities, we are also saying how we think they ought to be treated. That is, the same assumptions that in practice attribute such capacities to other humans also gauge their moral standing.

I suggest the same is and ought to be true in the case of our dealings with technological artifacts. From the perspective proposed here, accordingly, assessing such capacities in machines is, among other things, inescapably a moral issue. To assume that a machine has a consciousness and feelings akin to our own is at the same time to posit that it has moral standing, that it deserves to be treated in ways akin to how we treat fellow humans. Therefore, if in assessing whether a machine is conscious, has feelings, etc. we are to use our interactions with human beings as a guide (as the Cook Test suggests), then criteria that leave us unable to assess whether our interactions with such machines are morally responsible also leave us unable fully to assess whether they are effective. Put another way, if we cannot account for how we assess their moral standing (if we are asked, even implicitly, to "pay no attention to" this), any assumptions we may make about their capacities would be unlike those we make about people and, thus, would fail the Cook Test. The perspectives of the Tricksters and the Roommates, like most of the Great Debate, provide in themselves no clear foothold for making moral assessments of our interactions with the technologies they otherwise consider in great depth and detail. Consequently, their claims to assess what capacities computers do and do not have are, in the terms suggested here, unrealistic, incomplete and inconclusive.

Moreover, if we are looking at the potential of making machines with capacities that in ourselves contribute to our moral standing, we need to be able to assess not only how those machines ought to be treated, but also whether bringing them into existence (or even attempting to) is itself morally justified. This includes being able to address such questions as: Is it acceptable to make a technological artifact that can think and feel as we do, but which we can enslave, unplug, or reprogram? Would we be justified in developing technologies that would make it possible to produce ten-thousand sentient robots that are indistinguishable from one another, each one of which could consider itself to be the true one? What moral sense do we make of our producing conscious, feeling entities that are potentially immortal? How ought we to assess the moral or political demands of a robot or a class of robots? Is it responsible professional practice for computer scientists to work toward the development of sentient machines while leaving moral assessment of their work to others or to the future?

Parallel questions that concern bringing humans into the world are commonly treated as moral ones. (They are not questions always asked by everyone who procreates; but they are always questions that *can* be asked and often need to be assessed.) This is why philosophers and the

general public alike question the use of fertility drugs that can result in multiple births; or question conceiving a child wholly or partly to use his or her bone marrow to treat a disease in an older sibling; or question public policies that encourage or discourage a particular family size. In short, moral questions with respect to "building other human beings" are never inappropriate: in any given case the answer may be "it is morally justified" or "it's unacceptable" or "I don't know," but the questions themselves are reasonable and responsible ones to ask.

Since questions of this kind are moral ones when applied to humans, it is essential to be able to address parallel questions with respect to machines envisioned to have the sort of capacities that make ourselves beings with moral standing. In the spirit of the Cook Test, we must be able to assess the moral aspects of what we may assume about such machines precisely because we treat this as essential in the case of humans.

As for myself, I have yet to encounter a gadget that I felt I must assume has thoughts or feelings in order for me to interact with it in an effective and responsible way. Nor can I imagine what such an artifact might be—except in the question-begging sense noted above. It may make for better use of a technology to treat it *as if* it were intelligent (or cranky or stupid), but in doing so I am not obliged in the slightest, not even by an elaborately constructed "polite convention," to assume that the device *in fact* possesses those characteristics.

Ultimately, it needs to be more commonly acknowledged that there is nothing we know now about ourselves or about technology that requires us to conclude that an artifact with the kinds of human capacities considered here could never be made. Nor is there anything that enables us to conclude that it could. These brute facts stand firmly at the centre of the Great Debate, although almost always behind curtains of such cunning design that they go quite unnoticed. In any event, we would do well to remember that being able to draw distinctions between the given and the made, between the model and the modeled, and between ourselves and our artifacts has always been essential to the effective and responsible use of our technologies, and perhaps will—or should—always remain so. Accordingly, if one day, with all known curtains drawn open, it proves *possible* to make a thinking, conscious and feeling artifact, I am at a loss to imagine any criteria at all that would enable us to find this a *responsible* thing to do.

For the time being at least, as the Friends of Toto would remind us, if responsible progress is to be made in the Great Debate—or in the technological projects with which it is associated—we need to give serious attention to the *various* criteria relevant to its claims and speculations. This includes being able to assess both the possibility and the acceptability of the things we make and propose to make. It behoves us in all this to look for where curtains might be drawn closed by the Great Debate's various camps—or by ourselves.

References

Agre, P.E. 1997. *Computation and Human Experience*. Cambridge: Cambridge University Press.

Anderson, A.R. (ed.). 1964. *Minds and Machines*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bolter, J.D.. 1984. *Turing's Man: Western Culture in the Computer Age*. Chapel Hill: The University of North Carolina Press.

Clark, A. 2001. *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford, UK: Oxford University Press.

Collins, H.M. 1990. *Artificial Experts: Social Knowledge and Intelligent Machines*. Cambridge, MA: MIT Press.

Cook, S. D.N. 2008. "Design and Responsibility: The Interdependence of Natural, Artifactual, and Human Systems." in: *Philosophy and Design: from Engineering to Architecture*. Vermaas, Pieter E., Kroes, Peter, Light, Andrew, and Moore, Steven A. (eds). Dordrecht: Springer.

Dreyfus, H.L. 1979. *What Computers Can't Do*. New York: Harper Colophon Books.

Feigenbaum, D.A., and J. Feldman (eds). 1995. *Computers and Thought*. Menlo Park: AAAI Press/ The MIT Press.

George, F. H. 1979. *Philosophical Foundations of Cybernetics*. Tunbridge Wells, Kent UK: Abacus Press.
Graubard, S.R. 1988. *The Artificial Intelligence Debate: False Starts, Real Foundations*. Cambridge, MA: MIT Press.
Kurzweil, R. 1999. *The Age of Spiritual Machines*. New York, NY: Viking.
Penrose, R. 1989. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York: Penguin Books.
Raphael, B. 1976. *The Thinking Computer: Mind Inside Matter.* San Francisco: W. H. Freeman and Company.
Sayre, K.M. 1969. *Consciousness: A Philosophic Study of Minds and Machines*. New York: Random House.
Searle, J.R. 1980. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences* 3, 417-457.
Turing, A.M.1950. "Computing Machinery and Intelligence." *Mind* Vol. LIX No. 236, 433-460.

Endnotes

1 See Cook 2008.
2 The debate itself has been plotted at Stanford, and found to contain thousands of strands.
3 There are, of course, many antecedents to the modern debate, including, for example, observations by Descartes, Leibnitz, and Hobbes. And many contemporary figures have contributed a variety of important and sustained strands, including Hubert Dreyfus, Jerry Fodor, Paul and Patricia Churchland, Ned Block, and Ray Kurzweil, among others. Broader considerations of key issues can be found in such works as: Agre (1997); Anderson (1964); Bolter (1984); Collins (1990); Feigenbaum and Feldman (1995); George (1979); Graugard (1988); Penrose (1989); Raphael (1976); and, Sayre (1969).
4 In a similar vein, Searle and others have noted that there is a difference between simulating something and duplicating it.
5 There are many uses of the term "simulacrum." I focus here on the sense that seems most germane to the topic.
6 If such a computational simulacrum is ever created, I propose it be named "Zeno."
7 See, for example, Searle 1980, p. 422.
8 See, again, Searle 1980, p. 417 and 422
9 See, for example, Clark 2001, p.36.