The New Atlantis
A Journal of Technology & Society

Mind Games

Charles T. Rubin

The field of artificial intelligence poses special problems for how we think about the relationship between information, matter, and life. One premise of AI is that we can design and arrange matter into machines that can replicate one of the most distinctively human aspects of life, intelligence, by conceiving of it as a form of information processing. As AI has moved from a subject of science fiction to reality, increasing attention is being paid to its practical aspects, including the ethical and political questions it raises. Consider the report on AI published by the White House during the final months of the Obama administration. The report, called *Preparing for the Future of Artificial Intelligence*, begins in this way:

> Artificial Intelligence (AI) has the potential to help address some of the biggest challenges that society faces. Smart vehicles may save hundreds of thousands of lives every year worldwide, and increase mobility for the elderly and those with disabilities. Smart buildings may save energy and reduce carbon emissions. Precision medicine may extend life and increase quality of life. Smarter government may serve citizens more quickly and precisely, better protect those at risk, and save money. AI-enhanced education may help teachers give every child an education that opens doors to a secure and fulfilling life. These are just a few of the potential benefits if the technology is developed with an eye to its benefits and with careful consideration of its risks and challenges.

Notably absent from this list, and indeed from the report as a whole, are artificially intelligent humanoid robots that might serve as companions, caregivers, or sexual partners. Yet creating such AI devices, which would arguably have more need of a human-like form than other machines, is a possibility being pursued by many roboticists. Japan has become famous (but is hardly alone) for developing caregiver robots to deal with the anticipated deficit of people to look after an aging population. The real-world project to develop emotionally supportive robots is sufficiently well established to have generated its first apostate, Sherry Turkle. The prospect of sex-bots, much feted by both popular and academic futurists, is

***Charles T. Rubin**, a* New Atlantis *contributing editor, is an associate 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).*

now reportedly on its way to becoming a viable commercial reality. Actual results from roboticists so far often reveal a gap between the sophistication of the underlying technologies and the crudeness of the result, and between the sensationalistic claims of the headlines and the actual visible achievements. But work is ongoing.

In popular culture, of course, the notion of humanoid robots in our midst is very firmly established and has been for some time. It is currently being explored in the HBO series *Westworld*, a more sophisticated version of the 1973 Michael Crichton movie of the same name. The idea of a caregiver robot is sufficiently in the popular mind that the French chapter of the St. Vincent DePaul Society recently produced a five-minute advertisement for volunteer human companions premised on the insufficiency of robotic replacements. At least one recent movie (*Robot and Frank*) and one TV show (*Humans*) are premised on the prevalence of sophisticated helper robots in a not very distant future.

These emotionally rich applications of artificial intelligence would not necessarily push the boundaries of efforts to model as closely as possible both human physical and intellectual capacities. After all, people already engage in sex acts with inanimate objects and dolls; many of the emotionally supportive robots being created model human-animal interactions more than human-human interactions; and a nurse robot would not have to look like a human nurse to take a temperature or give an injection. Indeed, there are those who think it is foolish or unimaginative to take either human intelligence or the human body seriously when developing artificial intelligences that are, after all, to be used to overcome the weaknesses of the human mind and body. A self-driving car, for example, does not need a robot taxi driver behind the wheel. Yet our imaginations also push in the opposite direction, and not unreasonably. It is our particular form of embodiment that allows us to perform many functions that that same embodiment calls for. Our bodies and minds as they are allow us to use the tools and play the many assistive roles that human beings require because we are so embodied and minded. In addition, the familiar form of our embodiment provides the potential for being comforting and pleasurable in and of itself. When it comes to devices designed for sexual acts, for example, imitation of the embodied form of a human being presents certain obvious advantages.

Ex Machina, the 2015 film written and directed by Alex Garland, is all about what it would mean to create a robot that both looks and acts like a human being. Where the recent Obama administration report on AI contextualizes it within a framework of technological primacy in disembodied

expert systems that promote economic and social efficiency, Garland instead places the quest for AI within the longstanding human aspiration, reflected in stories ranging from the myth of Pygmalion to *Frankenstein*, to replace the natural procreation of new human beings with their technological manufacture. So doing allows him to dissect the motives behind the creation of such a being, the manner in which we would judge to what extent it had achieved human-like intelligence, and the dangers of succeeding at that goal.

Why Robot?

As *Ex Machina* begins, geeky programmer Caleb (Domhnall Gleeson) finds that he has won a trip to tech-wizard Nathan's (a thuggish Oscar Isaac) vast, remote mountain estate, a combination bunker, laboratory, and modernist pleasure-pad. Caleb at first does not know the purpose of the visit, only that he has won time with his boss and company founder in a company-wide lottery. (He later finds out he was not randomly chosen.) Only after he has signed the necessary non-disclosure agreement does Nathan reveal that Caleb is there in order to participate in a week-long test of Nathan's latest invention: an artificially intelligent robot called Ava (Alicia Vikander).

Nathan seems to have embarked on this project to create a human-like robot for many reasons. First, he may be doing it simply because he can. His position at the top of his company, Bluebook—a sort of mashup of Google and Facebook—gives him unique access, as he points out, to a huge world of data. (Some of it is acquired, he admits, by clandestine monitoring of cell phones.) That data, he believes, is the best foundation upon which to build artificial intelligence that goes beyond manipulating facts and figures to replicate "how people are thinking." He uses hacked cell phone cameras and microphones, for example, to "train" an AI to get facial expressions right. As a programmer, he sees all of this data as a great opportunity. AI is the brass ring of his profession, and he thinks he can grab it; as far as he is concerned, achieving it is only a matter of time.

Second, despite, or perhaps because of, his fantastic worldly success, Nathan does not like the world very much, and has spent vast sums of money attempting to escape from it. His estate somewhere in a mountain fastness takes hours to traverse by helicopter. He plainly resents the fact that "no matter how rich you get, s— goes wrong, you can't insulate yourself from it." In the face of this isolation from and impatience with human

beings, Nathan's AI research has the benefit of allowing him to create a cook/servant/sexual partner named Kyoko (Sonoya Mizuno), the next-to-last in a longish line of developmental iterations of attractive young woman robots, whose bodies he keeps, Bluebeard-like, in his bedroom.

Next, in an only partially comic moment, Nathan reveals that he would like to be a god. Caleb, overwhelmed by the prospect of what Nathan says he has done and by the formidable presence of his scary boss, suggests, in admittedly clichéd fashion, that the creation of "a conscious machine" would imply god-like power. We find the next day that what Nathan claims to remember, with a mixture of amusement and pride, is Caleb saying he would *be* a god.

What is "godlike" about the power to create a human-like AI—what would prompt Caleb to use that cliché? Success could hardly make Nathan more famous or wealthy; he has those things already and they have either driven him from the world or facilitated his departure. And Nathan already has a kind of omniscient power of surveillance. Given Caleb's severely limited knowledge of Ava at the time he makes the claim—he has only had one session encountering her—he would seem to associate godlikeness with the power to create a sentient being by non-biological means. But is his act of creation entirely godlike? Nathan determines the particular appearances of his creations, but he does not exercise godlike power over their overall form—rather, he models them on the form of attractive young women, a paradigm he did not create. The model for his robots' human-like behavior and thought is another given—one that Nathan takes from his heaps of surreptitiously collected data of human interactions with their devices—and so the minds, too, of these robots are not the sole product of his creative will. Finally, even with the technological breakthroughs that allow him to create his robots, Nathan is standing on the shoulders of others. In these respects he is no more godlike than any parent (he alludes to himself, once, as being like Ava's dad), or perhaps to a Prometheus, whose contribution to the human race was not a creation but a theft.

Yet Caleb may still be on to something. For if Nathan does not have the power to create entirely on his own, he certainly has the power to destroy with impunity, as the fate of the earlier robot models in his bedroom implies. But more than that, he thinks his work ultimately has destructive consequences for the fate of present intelligent life on earth. He expects that he is ushering in a new age, the Singularity, which he defines as the coming replacement of humans by superior forms of intelligence: "One day the AIs are gonna look back on us the same way we look at fossil skeletons in the plains of Africa: an upright ape, living in

dust, with crude language and tools, all set for extinction." His godlike power is largely, then, a power to destroy; he sees his creation of Ava in Oppenheimeresque terms. Having explained to Caleb his part in bringing on the Singularity, which he thinks may arrive with the model after Ava, he does not demur when Caleb quotes Oppenheimer's reaction to the atom bomb: "I am become Death, the destroyer of worlds."

So Nathan's motives range from the selfish to the misanthropic. He does not claim to be charitably or philanthropically motivated, to be thinking of his invention as something that might aid the lonely, the disabled, or the ill. He makes no representations that what he is doing will enrich the human experience—quite to the contrary. He does not mention the boons of technological innovation or economic efficiency even to his own company, let alone to mankind. His most "godlike" power turns out to be the power to destroy. This motive in particular, as we are about to see, colors his understanding of how to prove that he has successfully created artificial intelligence.

How Would We Recognize Artificial Intelligence?

Even though Nathan's nominal purpose for asking Caleb to join him is so that Caleb can be part of a "Turing Test" that will determine if indeed Ava is an artificial intelligence, Caleb's actual assignment has very little to do with Alan Turing's proposal, and as the plot develops the divergence widens. Nathan has his own ideas about what would constitute artificial intelligence or, as he sometimes says, consciousness. Since those differences go to the heart of the film, it will be helpful to remind ourselves briefly of how Turing's original version of the test, the "imitation game," was supposed to work and what Turing thought it could demonstrate.

In his original 1950 paper, Turing starts his discussion of the imitation game as if he were interested in the question "Can machines think?" However, to approach this question as if "the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used" would be to make it "difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd." Indeed, "The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion."

The imitation game effectively replaces *Can a machine think?* with a new question: *Can people be made to think a machine thinks?* The game Turing used for the test of artificial intelligence is modeled on a game

where a questioner communicates via terminal with two people, and on the basis of their answers to his questions decides which one is a man and which a woman. To adapt this game as a test of artificial intelligence, the man, as Turing specifies, is replaced by a machine. Turing asks, "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" He speculates that within fifty years (that is, by the year 2000) computers would "play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning."

Despite the fact that this famous essay is mostly about anticipated capacities of computers, notice that Turing reframed the question so we do not have to concern ourselves with what the machine is doing at all. He adopts the orientation of behaviorism, the then-growing school of psychological science that treats the mind (or in this case, the machine) as a "black box" and studies the perceivable results from perceivable stimuli as they arise from unknown "interior" processes. This approach avoids thorny questions like the relationship of mind to brain; it is a psychology that is indifferent to the existence or non-existence of the psyche. On the subject of thinking machines, what is worth considering under this behaviorist point of view is how people relate to the machine, and Turing expected that by the year 2000, much of the time people would be able to interact with machines as if they were interacting with human beings. He concludes that "at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."

Subsequent versions of the test drop the baseline of sexual differentiation and assume that the questioner is simply trying to determine which of multiple interlocutors is human and which not. In *Ex Machina*, Nathan modifies the test much further. For Caleb does not interact with Ava via a terminal, and there is no human interlocutor designated to serve as an experimental control. Rather, Caleb speaks directly with her from inside a small observation room, with thick glass between them. Ava's living quarters are a larger cage; she is not permitted access to him or to the rest of the facility. His first sight of Ava is in her most robotic instantiation, complete with transparent limbs and glowing abdomen. Her unclothed conformation is attractively female from the start, but only her face, hands and feet have human-like skin. The film's special effects and Vikander's acting do a wonderful job of making this combination deeply strange but hardly off-putting; Ava is fascinating and Caleb is quickly fascinated.

Caleb (Domhnall Gleeson) and Ava (Alicia Vikander) during their first session. Visible in the upper left are mysterious crack marks in the glass that separates them, as well as a camera Nathan uses to watch from another room.

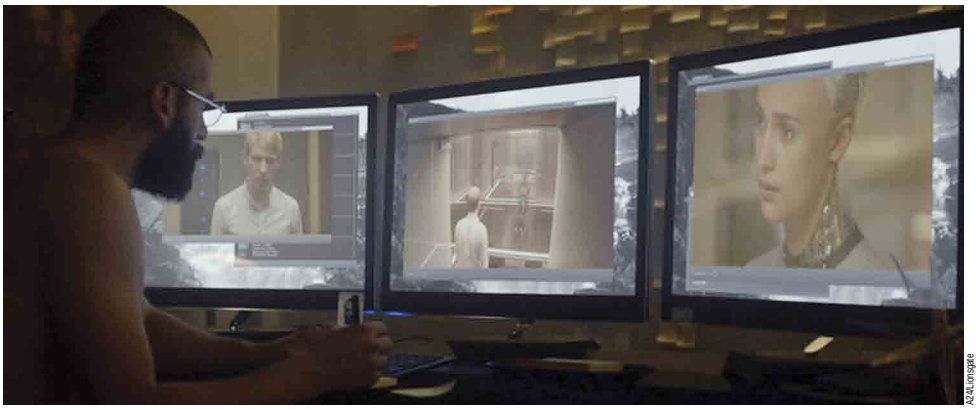
Caleb, therefore, knows from the start that the being he is interacting with is a machine and not a human, and so Nathan's test diverges sharply from Turing's in this respect. The reason for doing the test this way, Nathan claims early on, is to find whether Caleb is convinced she has consciousness *even knowing full well that she is a robot*: "If I hid Ava from you, so you just heard her voice, she would pass for human. The real test is to show you that she's a robot and then see if you still feel she has consciousness." The emphasis on consciousness is another difference from the Turing Test: unlike Nathan, Turing did not frame the imitation game as a test of consciousness. "I do not wish to give the impression that I think there is no mystery about consciousness," he wrote, "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." That is because for Turing the question is not, *Does the machine have consciousness?* but rather, in effect, *Do people think the machine has consciousness?* Thus, Nathan's alteration of the test is consistent with this orientation; he wants to know if Caleb feels Ava has consciousness even knowing that she is a machine.

Turing gave enough thought to the question of how a human-like form might influence the judgment of machine intelligence to remove deliberately that variable from his test. The imitation game, he thought, has "the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man." Even if engineers could produce material indistinguishable from human skin, there would be "little point in trying to make a 'thinking machine' more human by dressing it up in

such artificial flesh. The form in which we have set the problem reflects this fact in the condition which prevents the interrogator from seeing or touching the other competitors, or hearing their voices."

For Turing, the way in which skin and form may make a thinking machine "more human" is irrelevant, since he is concerned only with what is human with respect to "intellectual capacities." Caleb would seem to agree. He asks why Nathan did not simply present to him Ava's AI as "a gray box." He wonders if Nathan is trying to distract him by making Ava appear to be an attractive young woman, one who even seems to be flirting with him. Caleb's concern that Nathan wanted him "distracted" in a way that would produce a feeling in him that Ava is conscious is only made to seem more reasonable when it is finally revealed that her distracting appearance is anything but unintentional: she was designed with an eye to Caleb's preferences in women, which Nathan knows because his company gives him access to all of Caleb's digital communications, including his searches for pornography.

But on its face this situation is very strange. There was no necessity behind Nathan's choice to up the ante with respect to modifying the Turing Test and allowing Caleb to see Ava at all. If he really thought she could pass easily if invisible, but did not think an embodied Ava could pass without "distraction," why do the test in a new way in the first place? It hardly makes sense for Nathan to make up a test that he had to cheat on in order to pass if Ava could have passed a test he did not make up. As the movie progresses, and as Nathan's test departs yet further from Turing's, we find that suspicions on this point are justified.



The three lead characters of Ex Machina *are never together in one room. The only occasions they are seen together is when some of them appear on a screen, as when Nathan (Oscar Isaac, at left) watches Caleb and Ava's first session.*

Turing treats embodiment as somehow incidental to whether we think a machine thinks or not, or is conscious or not. For Nathan, the body is not so easily dismissed. He speculates, for example, that sexuality is an imperative that drives interaction. "Can you give an example of consciousness at any level, human or animal, that exists without a sexual dimension?" he asks. "What imperative does a gray box have to interact with another gray box? Can consciousness exist without interaction?" Such an interpretation of consciousness could be dismissed as merely a reflection of Nathan's own desire to create a robotic sexual partner, were it not for the fact that in the real world, most of the interactions that develop our ability to distinguish between conscious and unconscious (or less conscious) beings are interactions with bodies. The information conveyed by a terminal is just a small fraction of what we would normally use to call something a thinking or conscious being. Nathan's modifications of the Turing Test suggest a disagreement with Turing on the range of ways in which we perceive thinking or consciousness in the first place, an enrichment that would make Nathan the more comprehensive behaviorist, as Ava's behaviors, like human behaviors, extend beyond the communication via written or spoken language to encompass "body language."

This trans-Turing Test is not only evident in the interviews between Caleb and Ava, it is also strikingly illustrated in his interaction with Kyoko. We and Caleb alike eventually find out that she is one of Ava's predecessors. But when Caleb is introduced to her, Nathan tells him that there is no point in his trying to talk with her, as she speaks no English. Caleb, already uncomfortable with her apparently servant-like status, accepts this explanation at face value. He is likewise uncomfortable when at one point Kyoko seems to come on to him, and again when Nathan yells at her. But his discomfort throughout is premised on her humanity; apparently in this case her body language *alone* is enough to convince Caleb that he is dealing with a self-conscious, if somewhat strange, human being who simply does not speak his language. When Kyoko starts to unbutton her blouse in front of him, he does not respond as if he is thinking, *Ah, she must be a sex robot then!* but rather as if he is thinking, *The possibly stoned girlfriend/housekeeper of my boss thinks I want to have sex with her, when I really just want to know where Nathan is.*

As a result, Caleb is deeply shocked when he discovers that Kyoko is a robot, which he learns in the course of executing his plan to free Ava. Caleb takes the opportunity of Nathan's drunkenness to enter his bedroom-workroom to reprogram the building's security system. Kyoko watches as he also finds the evidence for previous robot models. She

stands naked before him and slowly peels some skin off her torso, and then off of her face. Thoroughly confused and distressed by his experiences, Caleb later examines his own body, and then confirms his flesh and blood status by using a razor to cut into his arm.

An AI that cannot communicate with words could never pass a Turing Test; it could not be a "thinking machine." Yet that conclusion hardly does justice to the deep disorientation that Caleb feels once he finds out he has been mistaken about Kyoko's status. Her example shows how judgments of consciousness are relational and contextual. Kyoko is a plausible non-English speaking, submissive, conscious human woman serving a powerful and likely abusive man. Were we to encounter a Kyoko in real life and be fooled as Caleb is, would it prompt disquiet about what the category of "conscious human woman" should be allowed to cover, or would we insist that we were just misled by a clever simulation? From Nathan's behaviorist point of view, at least, Caleb's complete acceptance of Kyoko as a fellow human being ought to be nearly as much a vindication of her "consciousness" as that which Ava achieves.

In the case of Kyoko, then, it is obvious that Nathan has so modified the Turing Test that he is not even trying to test what Turing wanted to test. Turing contributed to the blurred line between "thinking" and what would subsequently be called "artificial intelligence," and Nathan elides artificial intelligence into "consciousness." While he says he brings in Caleb to find out whether Caleb will treat Ava as if she were a conscious being, for all we know he is equally interested from the start in how Caleb treats Kyoko. To this end Nathan makes sure he is able to observe everything Caleb *does* with hidden and non-hidden cameras. Nathan stresses to Caleb that he wants to know how Caleb *feels* about Ava, and even more so he wants to make sure that Caleb is paying attention to the question of how Ava feels about him. At first Caleb is somewhat confused that Nathan is so interested in his feelings toward Ava; he plainly would rather talk about her programming. His limited ability to talk about his emotions is perfectly understandable under the circumstances: he is a guy, a programmer, he is talking to the big boss, and as his feelings for Ava develop, in her case at least he knows all along they are developing *for a robot.* But as he falls for Ava, his reason for reticence changes; to speak honesty of his feelings would be to reveal his desire to free Ava from her glass cage.

As the movie progresses, Caleb plainly is more and more convinced that Ava is a damsel in distress. At first, Caleb wondered if Ava feels anything at all. Perhaps she is interacting with him in accord with a highly sophisticated set of pre-programmed responses, and not experiencing her

Many of Caleb and Nathan's conversations to discuss Ava are over drinks, and Caleb eventually plans to take advantage of Nathan's drunkenness.

responses to him in the same way he experiences his responses to her. In other words, he wonders whether what she is feeling "inside" is the same as what he is feeling. Has she been programmed to flirt with him or does she really like him? When Caleb expresses such doubts, Nathan denies the legitimacy of the distinction. Drawing out the implications of his behavioristic materialism, he argues that Caleb's own desires are "a consequence of accumulated external stimuli that you probably didn't even register as they registered with you.... Of course you were programmed, by nature or nurture or both!" This is another of the unsettling thoughts that leads Caleb to the bloody investigation of his own humanity. But Nathan also tells Caleb, echoing *The Tempest*, that "For the record, Ava is not pretending to like you. And her flirting is not an algorithm designed to fake you out. You're the first man she's met that isn't me, and I'm like her dad, right? Can you blame her for getting a crush on you?"

There is a mirror-like quality to the test that Caleb is performing on Ava: The more Caleb sees Ava behaving as if she were fully aware of his consciousness or inner life, the more he acts as if she were experiencing what he would be experiencing in her place, and the more he believes, with her encouragement, that his feelings for her are reciprocated. Her mistrust of Nathan and her fears about her future move him even before Nathan explains to Caleb what her fate is to be. She is just one step in a developmental process for Nathan. He intends to

> download the mind. Unpack the data. Add the new routines I've been writing. To do that, you end up partially formatting, so the memories go. But the body survives.

Caleb comes to share Ava's belief that nobody should have the ability to shut her down in order to build the next iteration. He is convinced that she wants to be free, and that she deserves to be free. So he plots her escape, taking advantage of her ability to cause temporary power cuts in Nathan's facility.

Nature, Nurture, Design, and Freedom

When the time comes, however, it appears that Nathan was a step ahead of him, or at least so he claims on Caleb's last day. Caleb announces the big news that Ava has passed the test and that he and Nathan should celebrate by getting drunk one last time. Nathan's response is coy because he thinks he knows this is part of Caleb's plan to disable him and reprogram the security system to allow Ava to escape. He temporizes by baiting Caleb; did Caleb really come to grips with the question that had troubled him at the outset?

> Although I gotta say, I'm a bit surprised. I mean, did we ever get past the "chess problem," as you phrased it? As in: how do you know if a machine is expressing a real emotion or just simulating one? Does Ava actually like you, or not? Although now that I think about it, there is a third option—not whether she does or does not have the *capacity* to like you, but whether she's *pretending* to like you.…maybe if she thought of you as a means of escape.

Caleb cannot answer him intellectually, because (just as Nathan has encouraged) his judgment of Ava is emotional, not rational; she *feels* like a conscious being to him. And of course there is no way he can prove that Ava is not just using him as a human being might, and perhaps he is genuinely disturbed by the thought that she might be. But also, by this point he is simulating his own emotions for Nathan; he too is temporizing. As the hour scheduled for Caleb and Ava's escape plan approaches, Nathan claims that he knows the plan, but that an escape attempt is what he was hoping for all along. Showing Caleb how he encouraged Caleb's belief that Nathan treated Ava badly, he claims that the real test has been to see if Ava was sufficiently human to prompt Caleb—a "good kid" with a "moral compass"—to help her in precisely this way. "Ava was a rat in a maze. And I gave her one way out. To escape, she would have to use: self-awareness, imagination, manipulation, sexuality, empathy, and she did. Now if that isn't true AI, what the f— is?" Caleb feigns being bested, until the planned moment arrives, when he reveals that he does not have to get

Nathan drunk today to make it happen, having already reprogrammed the security system while Nathan was drunk the day before.

If Nathan all along wanted to see if Ava could get Caleb to help her escape, was he lying when he told Caleb that Ava was not designed to flirt with him? Nathan would distinguish between "programmed to flirt with Caleb" and "programmed to flirt in order to get what she wants." But that she was also programmed to want what she wants, including her freedom, is suggested by a video of one of Ava's predecessors, an earlier model called "Jade" that destroys her own arms beating on the glass of her cage in order to get out. The marks she left are still visible as Caleb interviews Ava. But recall that Nathan believes that people, no less than Ava, are "programmed." Therefore from his point of view, in creating Ava as a being he has programmed to want to be free, he is like one rat in a maze creating another rat in a maze.

So it is somewhat surprising to realize that Nathan, in the face of his self-understanding, would *like* to be a free man. In the context of the earlier conversation in which he claimed Caleb has been programmed by nature and nurture, he also says "The challenge is not to act automatically. It's to find an action that is not automatic, from painting to breathing to talking to f—ing to falling in love." Not acting automatically—that is, not acting out of the givens of nature and nurture—will obviously be difficult on his assumptions. That Nathan aspires to such fundamental freedom may help explain some of his own disdain for Ava; despite all the deeply human qualities he acknowledges she has exhibited, she is still to him a rat in a maze, a project involving a wall full of Post-It notes, because she is acting as he wishes her to act.

The movie's visuals support the notion that both Nathan and Ava are in a maze. Nathan's house is a maze-like series of traps, of doors that at any given moment may or may not freely open. Glass walls between rooms create disorienting reflections that conceal the actual physical layout in his bedroom/office, and the same effect is all the more obvious in Ava's quarters. A few of Nathan's rooms open to views of lush forest; Ava has a view into a room-sized terrarium with a few stunted trees in it.

The majesty of the natural environment his dwelling is situated in must be a mixed blessing for Nathan; when we see him outside he is likely to be using his punching bag. In the one scene where he and Caleb actually walk into the mountains, the camera pulls back to accentuate how small they are in the landscape; for Nathan the scenery must be a reminder of the power of the very givens that he is trying to overcome. How could Nathan know that he or his creations have overcome the given, the

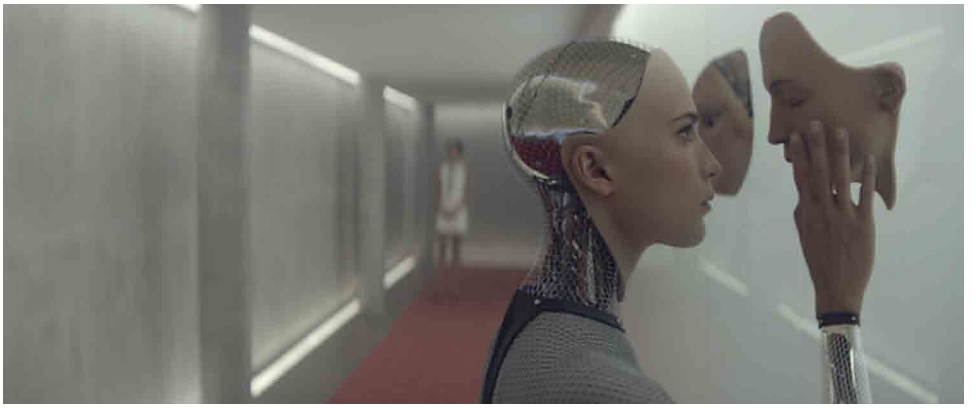
natural, the automatic? Strictly speaking it may never be possible, but one tempting option might be to do everything possible to negate as many of them as possible, to become, as we saw, "Death, the destroyer of worlds." To see such a drive to negation on the part of his creations would be, for Nathan, an indicator that they too are not bound by his givens. Such, at any rate, could be his answer to Ava's question, "Is it strange to have made something that hates you?"

Robots' Revenge

Apparently for Nathan one can be human and still operate "automatically"; an AI too, then, can be conscious in the sense he has been using the term and still programmed. Caleb as one kind of programmed being can decide that another kind of programmed being acts sufficiently like him to deserve his recognition and regard. Nathan's understanding of the kind of programming necessary to achieve this result is, as we have seen, far more sophisticated than Turing's, even if it is still programming. But now we see that there is yet a further "challenge": transcending the "automatic," the "given" of nature and nurture, altogether. The given is the maze for human intelligence and AI alike. For Nathan, Ava's proven ability to get Caleb to help her out of her maze is one more step towards the transhuman intelligence of the Singularity.

Writers on this topic call it "the Singularity" in order to stress precisely that this new form of intelligence will be incomprehensible to human beings. Nathan has some sense of this point in theory, but of course it is very difficult to know exactly what it would mean in practice, as he finds out. Neither he nor Caleb know that Ava and Kyoko have recently met (and, if their body language when we see them together after Ava's escape is to be trusted, developed some sort of intimate mutual understanding). We do not know the details of their meetings, but we observe the result of the unknown processes that would allow them to communicate: they have hatched their own plan within the framework of Caleb's escape plan.

When Nathan finds out that Caleb has outsmarted him and freed Ava, his first reaction is to give Caleb a disabling punch. His second is to arm himself with a dumbbell handle before he confronts Ava. The first action is perfectly comprehensible; he is very angry. The second may just follow from the first. Perhaps there is some sexual jealousy at work, and of course, anger is often not rational in its targets. But is there more? That Ava has told Nathan she hates him is surely relevant, but she said so as he was standing right next to her in her quarters—so far as we know he was



Out of her room for the first time, Ava sees a collection of masks in the hallway, including one that looks strikingly like her own face. Kyoko (Sonoya Mizuno) stands in the background. Moments later, the two robots will inaudibly conspire.

unarmed then. Is she at the same time strong enough to do him harm, *and* without any sort of kill switch or command? Does he know or intuit that she poses some threat free that she did not pose while caged?

I think the safest bet is that he is both displacing his anger with and jealousy of Caleb to Ava, *and* to some extent fearful of her. Perhaps a failsafe he thought he had programmed into her has already not worked—maybe she was not supposed to be able to exit her quarters, even if the door were open. Or perhaps he begins to see how the full potential of the capacity for human-like deception and manipulation that he has built into her may include a desire to be free that is more like his own will to negate the given than a simple rat-like wish to get out of a maze. When they confront each other, he claims he will not keep her locked up anymore, but she detects the lie (we know from her discussions with Caleb that she is a master at reading micro-expressions) and starts running toward him with obvious aggressive intent, a momentum-building run she would not have had space for in her quarters. He commands her to stop as if he expects she will obey, but she does not. So she is able to knock him down. As they struggle, she only briefly gets the better of him. He, in turn, uses the bar *after* she is more or less subdued; he makes a move to strike her and she shields herself with her forearm, which is shattered—a moment that seems to sober both of them.

But he has yet to experience his greatest achievement. For as Nathan walks backward dragging Ava by her feet to her room, Kyoko places herself so he runs into the chef's knife she is holding. Nathan smashes her in the head and she goes down, leaving the knife in his back. Ava removes it, and stabs him again in the chest. On Nathan's terms, these events are

a triumph. For how could the very limited Kyoko have plotted with Ava at all, and why would she do it, unless he had created even better than he knew? His dying words, "Okay… F—in' unreal… okay… oh… Ava" have an emotional range even in this context that could suggest a certain amazed pride. He has built AIs that embody his own willingness and ability to deceive, his own desire to be free in the form of a drive to negate the given. Perhaps he was even closer to the Singularity than he thought.

Meanwhile, Caleb comes to and sees Ava, who tells him to wait for her where he is. As he watches through glass she replaces her damaged arm with one from Nathan's earlier models, likewise takes from them skin, hair and a dress. Caleb sees her leave, but when he follows he finds he is locked in. Ignoring him, she meets the helicopter that was supposed to take Caleb home, as Caleb desperately tries to escape, fecklessly beating on glass with a stool, not unlike Nathan's earlier robot. Ava, he realizes, has deceived him about her feelings.

In the beautifully shot last scene of the movie, we see, upside down, Ava's shadow on a sidewalk. Earlier, when Caleb and Ava had discussed going on a date, she suggested as a venue a "busy pedestrian and traffic intersection in a city." Their date, Caleb joked, would consist of people-watching. He did not at that point know that her artificial intelligence is in large part the result of an unprecedented project of people-watching—Nathan's massive project of illicit digital snooping. Now, at the movie's end, we see Ava at just such an intersection before she disappears into the flow of humanity.

Throughout much of the movie, Caleb had been concerned that Ava did not "really" feel what he thought she was feeling, or what she represented herself as feeling. But such concerns are implicit violations of the behaviorist terms of Nathan's test. Caleb's decision to help her escape shows he puts these concerns about the reality of Ava's consciousness aside. That seems to have been a mistake. Nathan includes in Ava's programming the ability to manipulate, and although we can say that she treats Caleb in an inhumane way—it seems likely that, locked in as he is, he will starve to death—it is not exactly inhuman. Feigned romantic interest, betrayal, and even killing one's captors to escape confinement are all fairly ordinary (even if not common) human behaviors.

Failures of Self-Knowledge

Ex Machina presents us with a powerful picture of what it could mean, based on the behaviorist assumptions that undergird the classic Turing

Test, to achieve a human-like consciousness in a robot. But just as Nathan objects to the narrow range of behaviors that the classic test examines as relevant to intelligence, so the movie may be suggesting that we wonder even at the richer repertoire of "outputs" that Nathan introduces in order to achieve "consciousness." At the very least we can notice how his own selfish and destructive motives for creating AI are reflected in the behaviors he seeks to highlight as relevant to Ava's achievement of consciousness. Escaping her "programming" means recognizing the consciousness of others, and yet she uses her empathy to deceive and manipulate them.

We should hardly be surprised that Nathan has created Ava in his own image, or that having done so she should seek his life when given the chance. He can have all the data in the world on which to base her behaviors, but unless she can assimilate all that data herself he is still going to be the one to select what is relevant for what he wants her to be. When he thinks about what consciousness is he is thinking about himself, since on his assumptions that is the only consciousness he can truly know.

The cruel treatment of Caleb that follows from Nathan's understanding of consciousness may not put Ava outside the boundaries of human behavior. But stepping back from Nathan's assumptions, we have to wonder whether she would be any less human-like if she behaved as though she were grateful for what Caleb did for her—even if she did not really want to "be with" him. Indeed, precisely if the givens of nature and nurture are as unsatisfactory as Nathan believes them to be, couldn't gratitude be considered an example of a non-automatic ("gratuitous") act in the face of natural and culturally conditioned selfishness? Or, along different lines, if we are all trapped in the maze of our givens, why not an Ava who shows her freedom by behaving as if she were cultivating tranquil, contemplative acceptance of her fate?

It is not that Ava would have a human-like consciousness only if she were able to show us that she could have made all these choices, but rather that the more we consider such possibilities the more we can see that however much Nathan has expanded the behaviorist universe beyond Turing, he must still be narrowly selective. So long as she is useful to him, Nathan plainly wants Ava to be young and beautiful, so there is no reason to think he has designed her with any effort to duplicate the human life cycle. For one thing, although she looks like a young woman, she has no history. In her first interview with Caleb she offers to tell him her age: one. "One what?" Caleb asks, "One year or one day?" "One" she replies, almost interrupting him. Or again, we have no evidence that she would

enjoy eating or drinking, and without an organic metabolism, she would never age or put on weight.

In sum, even though she convinces Caleb that she fears her dissolution, Nathan has not given her the thousand little moments of mortality that form human consciousness. We might notice in contrast that in HBO's *Westworld* there is a robot whose increasingly dissonant experiences of its own world are leading it to something like "immortal longings." In comparison, Ava's concern with self-preservation seems purely prosaic. Aside from one big smile, her facial expressions as she departs Nathan's bunker seem more subdued than what we have previously seen; Ava is quietly satisfied that she has achieved her ends. She has done to Nathan pretty much what Nathan would have done to her, and she has arrived in the very world that he rejected.

There may not be many (if any) serious AI researchers for whom, as for Nathan and some transhumanists, the end of humanity is a feature, not a bug—although certainly that is true for some transhumanist *advocates* of AI research. Furthermore, most professional AI research programs, we are told, are not even attempting to create AI in the sense Turing wrote of, let alone in Nathan's more sophisticatedly "strong AI" version of it. Nevertheless, it remains true that it would certainly be useful for a subcategory of AI applications if it were possible to create a robot that could be mistaken for human, and some people are in fact trying to create such robots as replacements for human beings. *Ex Machina* alerts us to the possibility that even if Nathan's specific motives for creating AI are uncommon, the common motives (like profit, efficiency, uniformity, expertise) will no less condition what AI developers define success to be in this area. To the extent that developers are likely to take those motives for granted, or (like most of us) likely to be bad judges of their own motives, they may well, like Nathan, also be slow to see for themselves the risks of what they are doing. Developers may be unlikely to lose their lives as a result, but their work may still have bad consequences for third parties.

The movie also suggests that any effort to replicate human behaviors may likewise focus on a limited repertoire of behaviors whose manifestations follow from the motives of those creating the AI. The whole point of a humanoid robot is to overcome the emotional distance between man and machine, but if *Ex Machina* is correct, bridging that gap is as much or more a matter of need on the human side as it is of getting the behaviors right on the machine side. The danger here is that the very emotional or physical needs that are taken to call for robotic assistance or satisfaction will lead those receiving it to accept the partial versions of humanity they

are interacting with as human. The harm that could befall a vulnerable population does not, in order to be harm, have to look like what is done to Caleb. In a way, it is bad if in some manner the machine ends up disappointing the human user for an eventually revealed lack of humanity, and it is bad if the machine never disappoints because the very satisfaction gained from the relationship forecloses any desire for richer possibilities. If Caleb were to have gone off into the sunset with Ava, should we be happy for him, or sad that this emotionally fragile young man never grew up enough to find happiness with a real woman?

People hurt and help people all the time. A new route to creating "people" will not necessarily change this underlying situation for the better or the worse. But *Ex Machina* suggests that if we want to forestall bad outcomes, we need to pay attention to the motives of those creating humanoid robots and the selective repertoire of human behaviors that those motives are likely to produce. In the quest for better robots, we need to become better students of human character, and better at cultivating good human character. And perhaps a better understanding of ourselves would make us less likely to desire ever more human-like robots.