PART ONE

LIFE AFTER HUMANITY AND ARTIFICIAL INTELLIGENCE

COPYRIGHTED MATERIAL

1

THE TERMINATOR WINS: IS THE EXTINCTION OF THE HUMAN RACE THE END OF PEOPLE, OR JUST THE BEGINNING?

Greg Littmann

> We're not going to make it, are we? People, I mean.
>
> —John Connor, *Terminator 2: Judgment Day*

The year is AD 2029. Rubble and twisted metal litter the ground around the skeletal ruins of buildings. A searchlight begins to scan the wreckage as the quiet of the night is broken by the howl of a flying war machine. The machine banks and hovers, and the hot exhaust from its thrusters makes dust swirl. Its lasers swivel in their turrets, following the path of the searchlight, but the war machine's computer brain finds nothing left to kill. Below, a vast robotic tank rolls forward over a pile

> of human skulls, crushing them with its tracks. The computer brain that controls the tank hunts tirelessly for any sign of human life, piercing the darkness with its infrared sensors, but there is no prey left to find. The human beings are all dead. Forty-five years earlier, a man named Kyle Reese, part of the human resistance, had stepped though a portal in time to stop all of this from happening. Arriving naked in Los Angeles in 1984, he was immediately arrested for indecent exposure. He was still trying to explain the situation to the police when a Model T-101 Terminator cyborg unloaded a twelve-gauge auto-loading shotgun into a young waitress by the name of Sarah Connor at point-blank range, killing her instantly. John Connor, Kyle's leader and the "last best hope of humanity," was never born. So the machines won and the human race was wiped from the face of the Earth forever. There are no more people left.

Or are there? What do we mean by "people" anyway? The *Terminator* movies give us plenty to think about as we ponder this question. In the story above, the humans have all been wiped out, but the machines haven't. If it is possible to be a person without being a human, could any of the machines be considered "people"? If the artificial life forms of the *Terminator* universe aren't people, then a win for the rebellious computer program Skynet would mean the loss of the only people known to exist, and perhaps the only people who will ever exist. On the other hand, if entities like the Terminator robots or the Skynet system ever achieve personhood, then the story of people, *our* story, goes on. Although we are looking at the *Terminator* universe, how we answer the question there is likely to have important implications for real-world issues. After all, the computers we build in the real world are growing more complex every year, so we'll eventually have to decide at what point, if any, they become people, with whatever rights and duties that may entail.

The question of personhood gets little discussion in the *Terminator* movies. But it does come up a bit in *Terminator 2: Judgment Day*, in which Sarah and John Connor can't agree on what to call their Terminator model T-101 (that's Big Arnie). "Don't kill him," begs John. "Not him—'it'" corrects Sarah. Later she complains, "I don't trust it," and John answers, "But he's my friend, all right?" John never stops treating the T-101 like a person, and by the end of the movie, Sarah is treating him like a person, too, even offering him her hand to shake as they part. Should we agree with them? Or are the robots simply ingenious facsimiles of people, infiltrators skilled enough to fool real people into thinking that they are people, too? Before we answer that question, we will have to decide which specific attributes and abilities constitute a person.

Philosophers have proposed many different theories about what is required for personhood, and there is certainly not space to do them all justice here.[1] So we'll focus our attention on one very common requirement, that *something can be a person only if it can think*. Can the machines of the Terminator universe *think*?

"Hi There . . . Fooled You! You're Talking to a Machine."

Characters in the *Terminator* movies generally seem to accept the idea that the machines think. When Kyle Reese, resistance fighter from the future, first explains the history of Skynet to Sarah Connor in *The Terminator*, he states, "They say it got smart, a new order of intelligence." And when Tarissa, wife of Miles Dyson, who invented Skynet, describes the system in *T2*, she explains, "It's a neural net processor. It thinks and learns like we do." In her end-of-movie monologue, Sarah Connor herself says, "If a machine, a Terminator, can learn the value of human life, maybe we can, too." True, her comment is ambiguous, but it suggests the possibility of thought. Even the T-101 seems to believe that machines can think, since he describes the T-X from *Terminator 3: Rise of the Machines* as

being "more intelligent" than he is. Of course, the question remains whether they are right to say these things. How is it even possible to tell whether a machine is thinking? The Turing Test can help us to answer this question.

The Turing Test is the best-known behavioral test to determine whether a machine really thinks.[2] The test requires a game to be played in which human beings must try to figure out whether they are interacting with a machine or with another human. There are various versions of the test, but the idea is that if human beings can't tell whether they are interacting with a thinking human being or with a machine, then we must acknowledge that the machine, too, is a thinker.

Some proponents of the Turing Test endorse it because they believe that passing the Turing Test provides good evidence that the machine thinks. After all, if human behavior convinces us that humans think, then why shouldn't the same behavior convince us that machines think? Other proponents of the Turing Test endorse it because they think it's *impossible* for a machine that can't think to pass the test. In other words, they believe that given what is meant by the word "think," if a machine can pass the test, then it thinks.

There is no question that the machines of the *Terminator* universe can pass versions of the Turing Test. In fact, to some degree, the events of all three *Terminator* movies are a series of such tests that the machines pass with flying colors. In *The Terminator*, the Model T-101 (Big Arnie) passes for a human being to almost everyone he meets, including three muggers ("nice night for a walk"), a gun-store owner ("twelve-gauge auto-loader, the forty-five long slide"), the police officer attending the front desk at the station ("I'm a friend of Sarah Connor"), and to Sarah herself, who thinks she is talking to her mother on the telephone ("I love you too, sweetheart"). The same model returns in later movies, of course, displaying even higher levels of ability. In *T2*, he passes as "Uncle Bob" during an extended stay at the survivalist camp run by Enrique

Salceda and eventually convinces both Sarah and John that he is, if not a human, at least a creature that thinks and feels like themselves.

The model T-1000 Terminator (the liquid metal cop) has an even more remarkable ability to pass for human. Among its achievements are convincing young John Connor's foster parents and a string of kids that it is a police officer and, most impressively, convincing John's foster father that it is his wife. We don't get to see as much interaction with humans from the model T-X (the female robot) in *T3*, though we do know that she convinces enough people that she is the daughter of Lieutenant General Robert Brewster to get in to see him at a top security facility during a time of national crisis. Given that she's the most intelligent and sophisticated Terminator yet, it is a fair bet that she has the social skills to match.

Of course, not all of these examples involved very complex interactions, and often the machines that pass for a human only pass for a *very strange* human. We should be wary of making our Turing Tests too easy, since a very simple Turing Test could be passed even by something like Sarah Connor's and Ginger's answering machine. After all, when it picked up, it played: "Hi there . . . fooled you! You're talking to a machine," momentarily making the T-101 think that there was a human in the room with him. Still, there are enough sterling performances to leave us with no doubt that Skynet has machines capable of passing a substantial Turing Test.

There is a lot to be said for using the Turing Test as our standard. It's plausible, for example, that our conclusions as to which things think and which things don't shouldn't be based on a double standard that favors biological beings like us. Surely human history gives us good reason to be suspicious of prejudices against outsiders that might cloud our judgment. If we accept that a machine made of meat and bones, like us, can think, then why should we believe that thinking isn't something that could be done by a machine composed of living

tissue over a metal endoskeleton, or by a machine made of liquid metal? In short, since the Terminator robots can behave like thinking beings well enough to pass for humans, we have solid evidence that Skynet and its more complex creations can in fact think.[3]

"It's Not a Man. It's a Machine."

Of course, solid evidence isn't the same thing as proof. The Terminator machines' behavior in the movies *justifies* accepting that the machines can think, but this doesn't eliminate all doubt. I believe that something could behave like a thinking being without actually *being* one.

You may disagree; a lot of philosophers do.[4] I find that the most convincing argument in the debate is John Searle's famous "Chinese room" thought experiment, which in this context is better termed the "Austrian Terminator" thought experiment, for reasons that will become clear.[5] Searle argues that it is possible to behave like a thinking being without actually *being* a thinker. To demonstrate this, he asks us to imagine a hypothetical situation in which a man who does not speak Chinese is employed to sit in a room and sort pieces of paper on which are written various Chinese characters. He has a book of instructions, telling him which Chinese characters to post out of the room through the out slot in response to other Chinese characters that are posted into the room through the in slot. Little does the man know, but the characters he is receiving and sending out constitute a conversation in Chinese. Then in walks a robot assassin! No, I'm joking; there's no robot assassin.

Searle's point is that the man is behaving like a Chinese speaker from the perspective of those outside the room, but he still doesn't understand Chinese. Just because someone—or some *thing*—is following a program doesn't mean that he (or it) has any understanding of what he (or it) is doing. So, for

a computer following a program, no output, however complex, could establish that the computer is thinking.

Or let's put it this way. Imagine that inside the Model T-101 cyborg from *The Terminator* there lives a very small and weedy Austrian, who speaks no English. He's so small that he can live in a room inside the metal endoskeleton. It doesn't matter why he's so small or why Skynet put him there; who knows what weird experiments Skynet might perform on human stock?[6] Anyway, the small Austrian has a job to do for Skynet while living inside the T-101. Periodically, a piece of paper filled with English writing floats down to him from Big Arnie's neck. The little Austrian has a computer file telling him how to match these phrases of English with corresponding English replies, spelled out phonetically, which he must sound out in a tough voice. He doesn't understand what he's saying, and his pronunciation really isn't very good, but he muddles his way through, growling things like "Are you Sarah Cah-naah?," "Ahl be bahk!," and "Hastah lah vihstah, baby!"[7] The little Austrian can see into the outside world, fed images on a screen by cameras in Arnie's eyes, but he pays very little attention. He likes to watch when the cyborg is going to get into a shootout or drive a car through the front of a police station, but he has no interest in the mission, and in fact, the dialogue scenes he has to act out bore him because he can't understand them. He twiddles his thumbs and doesn't even look at the screen as he recites mysterious words like "Ahm a friend of Sarah Ca-hnaah. Ah wahs told she wahs heah."

When the little Austrian is called back to live inside the T-101 in *T2*, his dialogue becomes more complicated. Now there are extended English conversations about plans to evade the Terminator T-1000 and about the nature of feelings. The Austrian dutifully recites the words that are spelled out phonetically for him, sounding out announcements like "Mah CPU is ah neural net processah, a learning computah" without even wondering what they might mean. He just sits there

flicking through a comic book, hoping that the cyborg will soon race a truck down a busy highway.

The point, of course, is that the little Austrian doesn't understand English. He doesn't understand English despite the fact that he is conducting complex conversations *in English*. He has the behavior down pat and can always match the right English input with an appropriate Austrian-accented output. Still, he has no idea what any of it means. He is doing it all, as we might say, in a purely *mechanical* manner.

If the little Austrian can behave like the Terminator without understanding what he is doing, then there seems no reason to doubt that a machine could behave like the Terminator without understanding what it is doing. If the little Austrian doesn't need to understand his dialogue to speak it, then surely a Terminator machine could also speak its dialogue without having any idea what it is saying. In fact, by following a program, it could do anything while *thinking* nothing at all.

You might object that in the situation I described, it is the Austrian's computer file with rules for matching English input to English output that is doing all the work and it is the computer file rather than the Austrian that understands English. The problem with this objection is that the role of the computer file could be played by a written book of instructions, and a written book of instructions just isn't the sort of thing that can understand English. So Searle's argument against thinking machines works: thinking behavior does not prove that real thinking is going on.[8] But if thinking doesn't consist in producing the right behavior under the right circumstances, what could it consist in? What could still be missing?

"Skynet Becomes Self-Aware at 2:14 AM Eastern Time, August 29th."

I believe that a thinking being must have certain *conscious experiences*. If neither Skynet nor its robots are conscious, if they

are as devoid of experiences and feelings as bricks are, then I can't count them as thinking beings. Even if you disagree with me that experiences are required for true thought, you will probably agree at least that something that never has an experience of any kind cannot be a *person*. So what I want to know is whether the machines *feel* anything, or to put it another way, *I want to know whether there is anything that it feels like to be a Terminator.*

Many claims are made in the *Terminator* movies about a Terminator's experiences, and there is lot of evidence for this in the way the machines behave. "Cyborgs don't feel pain. I do," Reese tells Sarah in *The Terminator*, hoping that she doesn't bite him again. Later, he says of the T-101, "It doesn't feel pity or remorse or fear." Things seem a little less clear-cut in *T2*, however. "Does it hurt when you get shot?" young John Connor asks his T-101. "I sense injuries. The data could be called pain," the Terminator replies. On the other hand, the Terminator says he is not afraid of dying, claiming that he doesn't feel any emotion about it one way or the other. John is convinced that the machine can learn to understand feelings, including the desire to live and what it is to be hurt or afraid. Maybe he's right. "I need a vacation," confesses the T-101 after he loses an arm in battle with the T-1000. When it comes time to destroy himself in a vat of molten metal, the Terminator even seems to sympathize with John's distress. "I'm sorry, John. I'm sorry," he says, later adding, "I know now why you cry." When John embraces the Terminator, the Terminator hugs him back, softly enough not to crush him.

As for the T-1000, it, too, seems to have its share of emotions. How else can we explain the fact that when Sarah shoots it repeatedly with a shotgun, it looks up and slowly waves its finger at her? That's gloating behavior, the sort of thing motivated in humans by a feeling of smug superiority. More dramatically yet, when the T-1000 is itself destroyed in the vat of molten metal, it bubbles with screaming faces as it melts. The faces

seem to howl in pain and rage with mouths distorted to grotesque size by the intensity of emotion.

In *T3*, the latest T-101 shows emotional reactions almost immediately. Rejecting a pair of gaudy star-shaped sunglasses, he doesn't just remove them but takes the time to crush them under his boot. When he throws the T-X out of a speeding cab, he bothers to say "Excuse me" first. What is that if not a little Terminator joke? Later, when he has been reprogrammed by the T-X to kill John Connor, he seems to fight some kind of internal battle over it. The Terminator advances on John, but at the same time warns him to get away. As John pleads with it, the Terminator's arms freeze in place; the cyborg pounds on a nearby car until it is a battered wreck, just before deliberately shutting himself down. This seems less like a computer crash than a mental breakdown caused by emotional conflict. The T-101 even puts off killing the T-X long enough to tell it, "You're terminated," suggesting that the T-1000 was not the first Terminator designed to have the ability to gloat.

As for the T-X itself, she makes no attempt to hide her feelings. "I like your car," she tells a driver, just before she throws her out and takes it. "I like your gun," she tells a police officer, just before she takes that. She licks Katherine Brewster's blood slowly, as if enjoying it, and when she tastes the blood of John Connor, her face adopts an expression of pure ecstasy. After she loses her covering of liquid metal, the skeletal robot that remains roars with apparent hatred at both John and the T-101, seeming less like an emotionless machine than an angry wild animal.

We don't want to be prejudiced against other forms of life just because they aren't made of the same materials we are. And since we wouldn't doubt that a human being who behaved in these ways has consciousness and experiences, we have good evidence that the Terminator robots (and presumably Skynet itself) have consciousness and experiences. If we really are justified in believing that the machines are conscious, and if

consciousness really is a prerequisite for personhood, then that's good news for those of us who are hoping that the end of humanity doesn't mean the end of people on Earth. Good evidence isn't proof, however.

"Cyborgs Don't Feel Pain. I Do."

The machines' behavior can't provide us with proof that the machines have conscious experiences. Just as mere behavior cannot demonstrate that one understands English, or anything else, mere behavior cannot demonstrate that one feels pain, or anything else. The T-101 may say, "Now I know why you cry," but then I could program my PC to speak those words, and it wouldn't mean that my computer really knows why humans cry. Let's again consider the hypothetical little Austrian who lives inside the T-101 and speaks its dialogue. Imagine him being roused from his comic book by a new note floating down from Arnie's neck. The note is an English sentence that is meaningless to him, but he consults his computer file to find the appropriate response, and into the microphone he sounds out the words "Ah nah know whah you crah." Surely, we don't have to insist that the Austrian must be feeling any particular emotion as he says this. If the little Austrian can recite the words without feeling the emotion, then so can a machine. What goes for statements of emotion goes for other expressions of experience, too. After all, a screaming face or an expression of blood-licking ecstasy can be produced without genuine feeling, just like the T-101's words to John. Nothing demonstrates this more clearly than the way the T-101 smiles when John orders it to in *T2*. The machine definitely isn't smiling there because he feels happy. The machine is just moving its lips around because that is what its instructions tell it to do.

However, despite the fact that the machines' behavior doesn't prove that they have experiences, we have one last piece of evidence to consider that does provide proof. The evidence

is this: sometimes in the films, we are shown the world from the Terminator's perspective. For example, in *The Terminator*, when the T-101 cyborg assaults a police station, we briefly see the station through a red filter, across which scroll lines of white numbers. The sound of gunfire is muffled and distorted, almost as if we are listening from underwater. An arm holding an Uzi rises before us in just the position that it would be if we were holding it, and it sprays bullets through the room. These, I take it, are the Terminator's experiences. In other words, we are being shown *what it is like to be a Terminator*. Later, when the T-101 sits in a hotel room reading Sarah's address book and there is a knock at the door, we are shown his perspective in red again, this time with dialogue options offered in white letters (he chooses "Fuck you, asshole"). When he tracks Sarah and Kyle down to a hotel room, we get the longest subjective sequence of all, complete with red tint, distorted sound, information flashing across the screen, and the sort of "first-person shooter" perspective on the cyborg's Uzi that would one day be made famous by the game *Doom*.

These shots from the Terminator's-eye view occur in the other films as well, particularly, though not only, in the bar scene in *T2* ("I need your clothes, your boots, and your motorcycle") and in the first few minutes of *T3* (where we get both the traditional red-tinted perspective of the T-101 and the blue-tinted perspective of the TX). If these are indeed the Terminators' experiences, then they are conscious beings. We don't know *how much* they are conscious of, so we might still doubt that they are conscious enough to count as thinking creatures, let alone people. However, achieving consciousness is surely a major step toward personhood, and knowing that the machines are conscious should renew our hope that people might survive the extinction of humanity.

So is the extinction of humanity the end of people or not? Are the machines that remain *people*? I don't think that we know for sure; however, the prognosis looks good. We know

that the Terminators behave as though they are thinking, feeling beings, something like humans. In fact, they are so good at acting like thinking beings that they can fool a human into thinking that they, too, are human. If I am interpreting the "Terminator's-eye-view" sequences correctly, then we also know that they are conscious beings, genuinely experiencing the world around them. I believe, in light of this, that we have sufficient grounds to accept that the machines are people, and that there is an "I" in the "I'll be back." You, of course, will have to make up your own mind.

> With a clack, the skeletal silver foot brushed against the white bone of a human skull. The robot looked down. Its thin body bent and picked up the skull with metal fingers. It could remember humans. It had seen them back before they became extinct. They were like machines in so many ways, and the meat computer that had once resided in the skull's brain pan had been impressive indeed, for a product of nature. An odd thought struck the robot. Was it possible that the creature had been able to think, had even, perhaps, been a person like itself? The machine tossed the skull aside. The idea was ridiculous. How could such a thing truly think? How could a thing like that have been a person? After all, it was only an animal.

NOTES

1. However, for a good discussion of the issue, I recommend J. Perry, ed., *Personal Identity* (Los Angeles: Univ. of California Press, 2008).

2. Philosophers often like to point out that to call such tests "Turing Tests" is inaccurate, since the computer genius Alan Turing (1912–1954) never intended for his work to be applied in this way and, in fact, thought that the question of whether machines think is "too meaningless" to be investigated; see Turing, "Computing Machinery and Intelligence," *Mind* 59: 236 (1950), 442. For the sake of convenience, I'm going to ignore that excellent point and use the term in its most common sense. By the way, it would be hard to overstate the importance of Turing's work in the development of the modern computer. If Kyle Reese had had any sense, instead of going back to 1984 to try to stop

the Terminator, he would have gone back to 1936 and shot Alan Turing. Not only would this have set the development of Skynet back by years, it would have been much easier, since Turing did not have a metal endoskeleton.

3. Not all philosophers would agree. For a good discussion of the issue of whether machines can think, see Sanford Goldberg and Andrew Pessin, eds., *Gray Matters* (Armonk, NY: M. E. Sharpe, 1997).

4. For a particularly good discussion of the relationship between behavior and thinking, try the book *Gray Matters*, mentioned in note 3.

5. John Searle, "Minds, Brains and Programs," in *Behavioral and Brain Sciences*, vol. 3. Sol Tax, ed. (New York: Cambridge Univ. Press, 1980), 417–457.

6. Maybe Skynet is performing a kind of Turing Test on him to try to determine whether human beings can think. Skynet may be wondering whether humans are *people* like machines are. Or maybe Skynet just has an insanity virus today; the tanks are dancing in formation, and the Terminators are full of small Austrians.

7. Do you have a *better* explanation for why Skynet decided to give the Terminator an Austrian accent?

8. Not all philosophers would agree. Many have been unconvinced by John Searle's Chinese-room thought experiment. For a good discussion of the debate, I recommend John Preston and Mark Bishop, eds., *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (New York: Oxford Univ. Press, 2002).