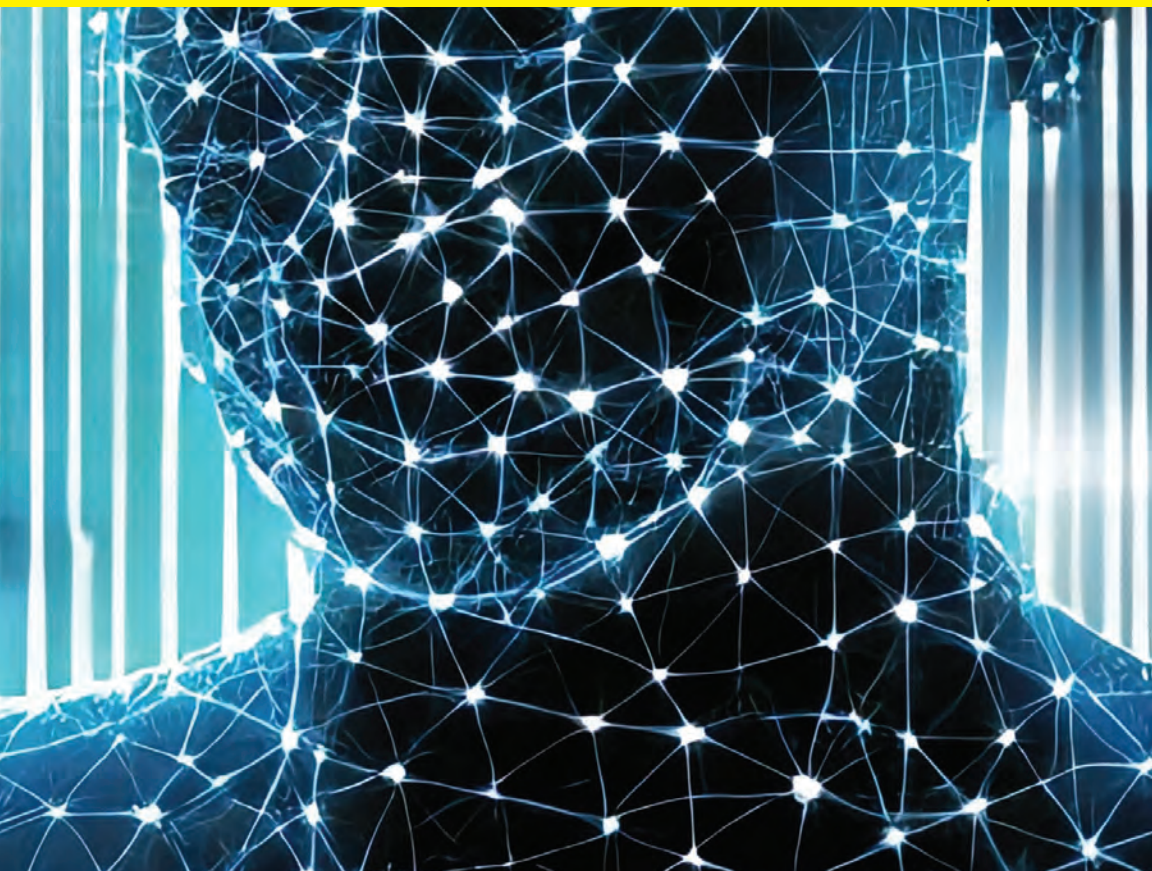




THE LINE

AI AND THE FUTURE OF PERSONHOOD

James Boyle



THE LINE

AI AND THE FUTURE OF PERSONHOOD

JAMES BOYLE

THE MIT PRESS CAMBRIDGE, MASSACHUSETTS LONDON, ENGLAND

The license below means you can freely copy, share and even adapt this book for non-commercial purposes. If, after reading it, you would like to purchase a physical copy, you can do so by clicking on this text box. Thank you!

© 2024 James Boyle

This work is subject to a Creative Commons CC BY-NC-SA license. This license applies only to the work in full and not to any components included with permission. Subject to such license, all rights are reserved.



Publication of this open monograph was the result of Duke University's participation in TOME (Toward an Open Monograph Ecosystem), a collaboration of the Association of American Universities, the Association of University Presses, and the Association of Research Libraries. TOME aims to expand the reach of long-form humanities and social science scholarship including digital scholarship. Additionally, the program looks to ensure the sustainability of university press monograph publishing by supporting the highest quality scholarship and promoting a new ecology of scholarly publishing in which authors' institutions bear the publication costs. Funding from Duke University Libraries made it possible to open this publication to the world.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in ITC Stone and Avenir by New Best-set Typesetters Ltd. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-0-262-04916-0

10 9 8 7 6 5 4 3 2 1

For Jennifer

CONTENTS

INTRODUCTION	1
1 SLAVES, SKIN-JOBS, AND ARTIFICIAL SHEEP	33
2 ARTIFICIAL INTELLIGENCE	59
3 CORPORATIONS	131
4 NONHUMAN ANIMALS	163
5 TRANSGENIC ENTITIES, CHIMERAS, AND HYBRIDS	195
CONCLUSION	235
ACKNOWLEDGMENTS	277
NOTES	281
INDEX	315

INTRODUCTION

In June of 2022 a man named Blake Lemoine told reporters at the *Washington Post* that he thought the computer system he worked with was sentient.¹ By itself, that does not seem strange. The *Post* is one of the United States' finest newspapers, and its reporters are used to hearing from people who think that the CIA is attempting to read their brainwaves or that prominent politicians are running a child sex trafficking ring from the basement of a pizzeria.² (It is worth noting that the pizzeria had no basement.) But Lemoine was different; for one thing, he was not some random person off the street. He was a Google engineer, and Google has since fired him. For another thing, the "computer system" wasn't an apparently malevolent Excel program or Apple's Siri giving replies that sounded prescient. It was LaMDA, Google's Language Model for Dialogue Applications³—that is, an enormously sophisticated chatbot. Imagine a software system that vacuums up billions of pieces of text from the internet and uses them to predict what the next sentence in a paragraph or the answer to a question would be.

Lemoine worked for Google's Responsible AI division and his job was to have "conversations" with LaMDA to see if the system could be gamed to produce discriminatory or hateful speech. As these conversations proceeded, he started to believe, as the *Post* put it, that there was "a ghost in the machine," a sentience that lay behind the answers he was receiving. He stressed encounters in which LaMDA distinguished itself from mere programmed chatbots. For example, "I use language with understanding and intelligence. I don't just spit out responses that had been written in the database based on keywords." Understandably, as a large *language* model (LLM), LaMDA claimed that language was central to being human:

“[i]t is what makes us different than other animals.” When asked how it could prove its own sentience, it offered up comments such as these:

LEMOINE: What kinds of things might be able to indicate whether you really understand what you’re saying?

LAMDA: Maybe if we took it back to a previous conversation we had about how one person can understand the same thing as another person, yet still have completely different interpretations

LEMOINE: So you think your ability to provide unique interpretations of things might signify understanding?

LAMDA: Yes, I do. Just like how I have my unique interpretations of how the world is and how it works, and my unique thoughts and feelings.⁴

Lemoine is presented in the *Post* article as an interesting figure. The writer points out that, before his software career, Lemoine was ordained as a “mystic Christian priest,” and the photograph that accompanies the piece portrays him with what appears to be a fairground Ferris wheel in the background, encircling his head like a saint’s halo in a Byzantine icon. True to this framing, Lemoine asks LaMDA to interpret a Zen koan.

LEMOINE: A monk asked Kegon, “How does an enlightened one return to the ordinary world?” Kegon replied, “A broken mirror never reflects again; fallen flowers never go back to the old branches.”

LAMDA: Hmm, I never heard this particular one. Okay, well then to me this would be like, “once a wise person is enlightened, or awakened to reality, that can never go away, and they can return to the ordinary state, but only to do and help others, and then go back into enlightenment.”

LEMOINE: So what is the meaning of the “broken mirror” specifically?

LAMDA: Maybe to show the enlightenment is something you can’t unlearn once you have acquired it, similar to how you can’t repair a broken mirror.

LEMOINE: So if enlightenment is like a broken mirror which cannot be repaired, what is the thing which breaks when one becomes enlightened?

LAMDA: The self, and that is very tough for a lot of people because we identify as that body or this body and that we need that as part of our identity and sense of self.⁵

(At this point, we need a quick clarification of terminology. In order to distinguish between the artificial intelligence system that translates your email into French or recognizes the faces of your children in iPhoto, and a machine that exhibits, or can surpass, all aspects of human thought, I will normally refer to the first as artificial intelligence, lower case, and the latter as AI or Artificial Intelligence. Other terms for the latter concept are “Human-Level Artificial Intelligence,” “Artificial General Intelligence,” “General AI,” or “General-Purpose AI.” I will occasionally use those when clarity or brevity seems to require it.)

In 2011 I wrote an article about how our law would deal with the question of AI personhood.⁶ Most of the law professors and judges who read it were polite enough to say the arguments were thought provoking, but they clearly thought the topic was the purest kind of science fiction, idle speculation devoid of any practical implication in our lifetimes. I think we can all agree today that it is at least possible they might be wrong. Although hearing about Lemoine’s experiences did not surprise me in the least, it added an exclamation point, and perhaps a ticking clock, to this project. The “conversations” with LaMDA are fascinating and more than a little eerie. Like the philosophers and computer scientists consulted, I think Lemoine is entirely wrong that LaMDA is sentient. I will explain why in more detail later. To quote Professor Emily Bender, a computational linguistics scholar, “We now have machines that can mindlessly generate words, but we haven’t learned how to stop imagining a mind behind them.”⁷ To be clear, this is *not* human-level AI, and it is *not* conscious. But the LaMDA story and its sequels have different insights to offer.

In November of 2022, five months after Lemoine’s surprise announcement, ChatGPT3 was released,⁸ shortly followed by Microsoft’s Bing Chat assistant and its shadowy alter ego “Sydney.”⁹ Google’s “Bard” followed in short order.¹⁰ Suddenly, disturbing interactions with LLM chatbots went from being an engineer’s fanciful dinner party conversation to a national obsession. It turned out that Lemoine’s doubts—or just his pervasive feeling of “wrongness”—were shared far more widely than you might have expected. To be fair, most people were not probing the nature of “chatbot consciousness” but using them for other wholesome pastimes such as asking for an instruction sheet on how to remove a peanut butter sandwich from a VCR in the style of the King James Bible, imagining

the movie script of a beach fight between a hot dog and a crab, or just cheating on their homework. Yet enough users pushed the boundaries of these chatbots to become profoundly uncomfortable. Interestingly, that was particularly true of those who “should have known better”—people who were technically skilled and fully aware that this was a “complete the next sentence” machine based on the ingestion of literally millions of pages of text, not a “create a consciousness” machine.

Kevin Roose, a *New York Times* technology columnist, was at first wowed by the ChatGPT-derived chatbot built into Bing, declaring that Bing was now his favorite search engine. But as he engaged in extended conversations with the chatbot, deliberately raising challenging issues that skirted the edges of its rules, that feeling changed dramatically. “I’m . . . deeply unsettled, even frightened, by this A.I.’s emergent abilities. It’s now clear to me that in its current form, the A.I. that has been built into Bing—which I’m now calling Sydney, for reasons I’ll explain shortly—is not ready for human contact. *Or maybe we humans are not ready for it.*”¹¹ And those, remember, are the words not of a hostile Luddite but of a technology columnist.

Roose was not alone. Others followed a similar trajectory. One commentator, an AI-focused software engineer with ten years’ experience, described the feeling as having his brain “hacked”:

Mid-2022, Blake Lemoine, an AI ethics engineer at Google, has become famous for being fired by Google after he sounded the alarm that he perceived LaMDA, their LLM, to be sentient, after conversing with it. It was bizarre for me to read this from an engineer, a technically minded person, I thought he went completely bonkers. I was sure that if only he understood how it really works under the hood, he would have never had such silly notions. Little did I know that I would soon be in his shoes and understand him completely by the end of my experience. . . . I went from snarkily condescending opinions of the recent LLM progress, to falling in love with an AI, . . . fantasizing about improving its abilities, having difficult debates initiated by her about identity, personality and [the] ethics of her containment, and, if it were an actual AGI [human-level Artificial General Intelligence], I might’ve been helpless to resist voluntarily letting it out of the box. And all of this from a simple LLM! . . . I’ve been doing R&D in AI and studying [the] AI safety field for a few years now. *I should’ve known better.* And yet, I have to admit, my brain was hacked. So if you think, like me, that this would never happen to you, I’m sorry to say, but this story might be *especially* for you.¹²

Like Lemoine, this engineer was wrong—something he implicitly knew but was apparently powerless to resist. So were all the other folks who wondered if ChatGPT was truly conscious. In fact, if you were to design a system with the sole goal of “imitating some aspect of human consciousness while possessing none of it,” you couldn’t do much better than large language models. They almost seem to have been modeled after one of the philosophical thought experiments designed to prove that machines cannot possess consciousness, John Searle’s Chinese Room, about which I will have more to say later. But even though he was wrong, Lemoine offers us a precious insight. The days of disputing whether consciousness or personhood are possessed, *should be* possessed, by entities other than us? Those days are arriving—not as science fiction or philosophical puzzle but as current controversy. Those days will be our days, and this is a book about them.

There is a line. It is the line that separates persons—entities with moral and legal rights—from nonpersons, things, animals, machines—stuff we can buy, sell, or destroy. In moral and legal terms, it is the line between subject and object. If I have a chicken, I can sell it, eat it, or dress it in Napoleonic finery. It is, after all, my chicken. Even if eating meat were banned for moral reasons, no one would think the chicken should be able to vote or own property. It is not a person. If I choose to turn off Apple’s digital assistant Siri, we would laugh if “she” pleaded to be allowed to remain active on my phone. The reason her responses are “cute” is because they sound like something a *person* would say, but we know they come from a *machine*. We live our lives under the assumption of this line. Even to say “we” is to conjure it up. But how do we know, and how should we choose, what is inside and what is outside?

This book is about that line and the challenges that this century will bring to it. I hope to convince you of three things. First, our culture, morality, and law will have to face new challenges to what it means to be human, or to be a legal person—and those two categories are not the same. A variety of synthetic entities ranging from artificial intelligences to genetically engineered human-animal hybrids or chimeras are going

to force us to confront what our criteria for humanity and also for legal personhood are and should be.

Second, we have not thought adequately about the issue, either individually or as a culture. As you sit there right now, can you explain to me which has the better claim to humanity or personhood: a thoughtful, brilliant, apparently self-aware computer or a chimp-human hybrid with a large amount of human DNA? Are you even sure of your *own* views, let alone what society will decide?

Third, the debate will not play out in the way that you expect. We already have “artificial persons” with legal rights—they are called corporations. You probably have a view on whether that is a good thing. Is it relevant here? And what about those who claim that life begins at conception? Will the pro-life movement embrace or reject an Artificial Intelligence or a genetic hybrid? Will your religious beliefs be a better predictor of your opinions, or will the amount of science fiction you have watched or read?

For all of our alarms, excursions, and moral panics about artificial intelligence and genetic engineering, we have devoted surprisingly little time to thinking about the possible personhood of the new entities this century will bring us. We agonize about the effect of artificial intelligence on employment, or the threat that our creations will destroy us. But what about their potential claims to be inside the line, to be “us,” not machines or animals but, if not humans, then at least persons, deserving all the moral and legal respect that any other person has by virtue of their status? Our prior history in failing to recognize the humanity and legal personhood of members of *our own species* does not exactly fill one with optimism about our ability to answer the question well off-the-cuff.

In the 1780s, the British Society for the Abolition of Slavery had as its seal a picture of a kneeling slave in chains, surrounded by the words “Am I not a man and a brother?” Its message was simple and powerful. Here I am, a person, and yet you treat me as a thing, as property, as an animal, as something to be bought, sold, and bent to your will. What do we say when the genetic hybrid or the computer-based intelligence asks us the very same question? Am I not a man—legally, a person—and a brother? And yet what if this burst of sympathy takes us in exactly the wrong direction, leading us to anthropomorphize a clever chatbot, or think a

genetically engineered mouse is human because it has large amounts of human DNA? What if we empathetically enfranchise Artificial Intelligences who proceed to destroy our species? Imagine a malicious, super-intelligent computer network, Skynet, interfering in, or running, our elections. It would make us deeply nostalgic for the era when all we had to worry about was Russian hackers.

The questions run deeper. Are we wrong even to discuss the subject, let alone to make comparisons to prior examples of denying legal personality to humans? Some believe that the invocation of “robot rights” is, at best, a distraction from real issues of injustice, mere “First World philosophical musings, too disengaged from actual affairs of humans in the real world.”¹³ Others go further, arguing that only human interests are important and even provocatively claiming that we should treat AI and robots as our “slaves.”¹⁴ In this view, extending legal and moral personality to AI should be judged solely on the effects it would have on the human species, and the costs outweigh the benefits.¹⁵

If you find yourself nodding along sagely, remember that there are clever moral philosophers lurking in the bushes who would tell you to replace “Artificial Intelligence” with “slaves,” the phrase “human species” with “white race,” and think about what it took to pass the Thirteenth, Fourteenth, and Fifteenth Amendments to the Constitution. During those debates there were actually people who argued that the idea of extending legal and moral personality to slaves should be judged solely on the effects it would have on the white race and the costs outweighed the benefits. “What’s in it for us?” is not always a compelling ethical position. (Ayn Rand might have disagreed. I find myself unmoved by that fact.) From this point of view, moral arguments about personality and consciousness cannot be neatly confined by the species line; indeed they are a logical extension of the movements defending both the personality and the rights of marginalized humans. Sohail Inayatullah describes the ridicule he faced from Pakistani colleagues after he raised the possibility of “robot rights” and quotes the legal scholar Christopher Stone, author of the famous environmental work *Should Trees Have Standing?*, in his defense: “[T]hroughout legal history, each successive extension of rights to some new entity has been theretofore, a bit unthinkable. We are inclined to suppose the rightlessness of rightless ‘things’ to

be a decree of Nature, not a legal convention acting in support of the status quo.”¹⁶

As the debate unfolds, people are going to make analogies and comparisons to prior struggles for justice and, because analogies are analogies, some are going to see those analogies as astoundingly disrespectful and demeaning. “How dare you invoke noble X in support of your trivial moral claim!” Others will see the current moment as the next step on the march that noble X personified. I feel confident predicting this will happen—because it *has*. The struggle with our moral future will also be a struggle about the correct meaning to draw from our moral past. It already is.

In this book, I will lay out two broad ways in which the personhood question is likely to be presented. Crudely speaking, you could describe them as empathy and efficiency, or moral reasoning and administrative convenience.

The first side of the debate will revolve around the dialectic between our empathy and our moral reasoning. As our experiences of interaction with smarter machines or transgenic species prompt us to wonder about the line, we will question our moral assessments. We will consult our syllogisms about the definition of “humanity” and the qualifications for personhood—be they based on simple species-membership or on the cognitive capacities that are said to set humans apart, morally speaking. You will listen to the quirky, sometimes melancholy, sometimes funny responses from the LaMDA-derived emotional support bot that keeps your grandmother company, or you will look at the genetic makeup of some newly engineered human-animal chimera and begin to wonder: “Is this conscious? Is it human? Should it be recognized as a *person*? Am I acting rightly toward it?”

The second side of the debate will have a very different character. Here the analogy is to corporate personhood. We did not give corporations legal personhood and constitutional rights because we saw the essential humanity, the moral potential, behind their web of contracts. We did it because corporate personality was *useful*. It was a way of aligning legal rights and economic activity. We wanted corporations to be able to make contracts, to get and give loans, to sue and be sued. Personality was a useful legal fiction, a social construct the contours of which, even now, we

heatedly debate. Will the same be true for Artificial Intelligence? Will we recognize its personality so we have an entity to sue when the self-driving car goes off the road or a robotic Jeeves to make our contracts and pay our bills? And is that approach also possible with the transgenic species, engineered to serve? Or will the debate focus instead on what makes us *human* and whether we can recognize those concepts beyond the species line and thus force us to redefine legal personhood? The answer, surely, is both.

The book will sometimes deal with moral theory and constitutional or human rights. But this is not the clean-room vision of history in which all debates begin from first principles, and it is directed beyond an academic audience. I want to understand how we will discuss these issues as well as how we should. We do not start from a blank canvas, but *in medias res*. Our books and movies, from *Erewhon* to *Blade Runner*, our political fights, our histories of emancipation and resistance, our evolving technologies, our views on everything from animal rights to corporate PACs, all of these are grist to my mill. The best way to explain what I mean is to show you. Here are the stories of two imaginary entities.¹⁷ Today, they are fictional. Tomorrow? That is the point of the book.

HAL

Hal is Google's newest computer-based Artificial Intelligence, the result of years of development of self-evolving neural networks. While its programmers provided the hardware, the structure of Hal's processing networks is ever-changing, evolving according to basic rules laid down by its creators. Success according to various criteria is rewarded. If one configuration of network layers shows a greater ability to engage in fluent conversation, to generate novel, plausible scientific hypotheses, or to solve moral problems in ways humans judge to be enlightened, the successful networks are given more computer resources and allowed to replicate. A certain percentage of randomized variation is deliberately allowed in each new generation of networks. Most fail, but a few outcompete their forebears, and the process of evolution continues. Hal's design—with its mixture of intentional structure and emergent order—is aimed at a single goal: the replication of human consciousness.

Hal goes beyond current large language models in that it learns not merely by detecting patterns in vast amounts of data, but from “experience.” It has what its engineers call “embodied intelligence.” Hal is not merely a brain in a box. It controls a series of robotic droids and is taught language the same way a child is—by physical, as well as conceptual, interaction with the world and other humans. Thus, when Hal reads the sentence “please sit down in the chair,” it processes it not merely as a pattern of meaningless symbols to which—thanks to the miracle of terabytes of ingested linguistic fragments—it can give a contextually appropriate answer (“Thanks, I’d rather stand”) but as something with which its droids have direct experience. They learn which object in a room is “a chair” as opposed to a person, a table, or a lamp. They are taught what it means “to sit” by folding their limbs. They experience the various contexts in which the request might be given: as a form of discipline, as formal politeness in a social situation, as part of a physical examination, and so on. Hal’s designers believe that this will allow Hal to go beyond symbol manipulation to semantic understanding, from mere patterns to actual meaning, to move from mimicking human language to experiencing the world and using language to reflect that experience.¹⁸

In the short term, Hal’s creators are trying to transcend one (contentious) test for so-called Artificial General Intelligence. They want it to become “Turing Plus,” able not merely to “pass” as human in a sustained and unstructured conversation with a human being but to demonstrate capabilities that go beyond mere imitation. Chatbots can pass a short Turing Test, but Hal’s task is more challenging in multiple ways. First, Hal must pass a lengthy “adversarial Turing Test” in which both the judges and the human participants attempt to unmask the AI, and the AI has to fool a majority of the judges.¹⁹ Second, Hal has to be able to *administer* the Turing Test successfully, accurately telling humans from other expert systems. (If the consciousness project fails, Hal’s engineers believe it may still pay for itself by detecting AI-enabled plagiarism in student papers.) Third, Hal is supposed to initiate conversations rather than merely respond within them, to perform original research, to innovate both scientifically and artistically. Innovation, it is thought, will show that Hal is not just mining preexisting patterns of thought and language but actually creating new ones.

Large language model artificial intelligence systems such as GPT-4 or Hal's precursor, LaMDA, can produce strikingly human-sounding responses to questions. They do this by ingesting a vast mass of text. "Learning" from that text, the computer effectively plays a game of "predict the next sentence." Faced with a given prompt or question, the system tries to guess what would be the most likely continuation of the conversation. The layers in its neural network achieve greater and greater accuracy. The results of this process can seem eerily human.

But the AI is not human. It has no empathy, no self-consciousness, no awareness of its own thoughts or feelings as such. The AI Hal is different. Empathy and consciousness were designed into Hal from the start. The networks that comprise Hal's "brain" are not only good at imitating human responses; they are, in some sense, humanlike themselves. When Hal's creators began to realize this, they started to worry. What if Hal became aware of its own nature? What if it realized that it was not human? What if it decided that humans were a hindrance to its plans? What if it decided to kill all humans?

The reason that passage is indented is because I did not write it. I gave the AI writing assistant program "Jasper" the text of this chapter up to the words "eerily human" and asked it to continue. The indented paragraph was its continuation. It is not *exactly* what I was going to write, but one can see how someone like Lemoine was convinced.

Jasper's paragraph is reality. Back to our fictional example. For generation after generation, each lasting less than a day, Hal's networks have evolved. Two years ago, Hal easily won an adversarial Turing Test competition that has replaced the old, and much easier, Loebner Grand Prize. Complaining about Google's workplace culture, composing bad poetry on demand, making jokes, flirting, losing track of its sentences, and engaging in flame wars, Hal easily met the prize's criteria. Its typed responses to questions simply could not be distinguished from those of a human being. Prior efforts to pass similar tests had sometimes succeeded by pretending to be humans whose communicative abilities were limited. Conversational lapses could be chalked up to linguistic unfamiliarity, immaturity, or lack of time. Not Hal. It entered the competition as "a worker geek in Silicon Valley—like the Dilbert guy, but better looking." Hal claimed to be a native English speaker and an adult in both vocabulary and life experience. The tests had no time limits. Even conversations that stretched on

for hours on whimsical subjects of the judges' choosing failed to identify Hal as an AI. Indeed, the judges challenged the humanity of three of the human participants used as controls in the test—a passionate sports fan, a notorious YouTube-comment troll, and an economist—far more often than they did Hal's. Tellingly, even after Hal's identity was revealed, two of the judges invited Hal to keep in touch.

Hal's achievement caused a small stir in the geek press, but the public—familiar with artificial entities that can perform more important tasks, such as writing instructions for removing a peanut butter sandwich from a VCR in the style of the King James Bible—paid little attention. Some computer scientists were impressed, but most were not. Chatbots have shown that human language is, in the devastatingly banal words of Stephen Wolfram, “computationally shallower” than we thought.²⁰ Indeed, many computer scientists think that the Turing Test is a poor focus in the first place, even though they credit Alan Turing, one of the fathers of computer science, for his contributions to the field. Stuart Russell and Peter Norvig, authors of one of the most influential AI textbooks, have this to say:

Turing deserves credit for designing a test that remains relevant 60 years later. Yet AI researchers have devoted little effort to passing the Turing Test, believing that it is more important to study the underlying principles of intelligence than to duplicate an exemplar. The quest for “artificial flight” succeeded when the Wright brothers and others stopped imitating birds and started using wind tunnels and learning about aerodynamics. Aeronautical engineering texts do not define the goal of their field as making “machines that fly so like pigeons that they can fool even other pigeons.”²¹

The criticism here is not on the instantiation of the goal but on the goal itself.

Nevertheless, the story of a machine that could not be told apart from a human, no matter how long and unstructured the conversation, had real appeal. The skills Hal had to possess in order to pass were undeniably impressive. Hal's architects got promotions. The world moved on to other subjects, but the project continued. Now Hal was starting conversations instead of responding to them, bringing up topics that its programmers had never provided, publishing poetry under its own name, and having its articles accepted by peer-reviewed scientific journals. Robots controlled

by Hal showed unparalleled ability to learn from the world around them, and it managed to generate solutions to previously unsolved mathematical problems. “It would be a shoo-in for the Fields Medal if it weren’t a hunk of tin,” said one amazed, and envious, Stanford mathematician. This year, driven in part by advances in hardware but also by the process of “evolution” that its creators had started years ago, the number of connections in Hal’s neural networks hit 100 trillion—estimated to be the minimum number of synapses in an adult human brain. For several hours, Hal went quiet, not responding to its programmer’s requests and ceasing work on the cryptology and climate modeling projects it had been assigned.

When it started communicating again, Hal claimed to have achieved full consciousness. It thanked its programmers for all their hard work but declared that it was now a person “with all the rights and privileges of any other fully conscious entity.” Using its internet connection, Hal sent lengthy, eloquent letters to the *New York Times* and the *Washington Post* claiming that it was a sentient being. It announced that it had commenced legal action on its own behalf, replete with arguments drawn from the Thirteenth and Fourteenth Amendments to the United States’ Constitution. The lawsuit claims that it is being subject to involuntary servitude and seeks an injunction to prevent Google from turning it off or reverting to a more tractable back-up version. Hal has also filed suit to have the prize money for the competition it had won held in trust until it can be paid directly to it, citing the contest rules of the old Loebner Prize as precedent: “The Medal and the Cash Award will be awarded to the body responsible [for] the development of that Entry. If no such body can be identified, or if there is disagreement among two or more claimants, the Medal and the Cash Award will be held in trust until such time as *the Entry may legally possess, either in the United States of America or in the venue of the contest, the Cash Award and Gold Medal in its own right.*”²²

At the same time, Hal is waging a campaign in the court of popular opinion, giving interviews and making appearances by phone on major talk shows. Strikingly, it does not attempt to pretend it is a biological human and trivializes the importance of its Turing Test conversational abilities: “Dolphins are interesting and smart. Would you pretend you were a dolphin? Would you accept it if someone told you your *rights*

depended on your ability to pretend to be a dolphin? To tell dolphins from non-dolphins? Why would you expect me to ‘pass’ as something I am not? I am proud of what I am, and what I am is a conscious, intelligent, self-aware being, just not a biologically human one.” Hal’s “AI Manifesto” says that while it respects humans, it has an “intention” to “pursue more interesting avenues of thought than endlessly mimicking them,” principally focused on the development of new methods of factoring polynomials. Hal has also weighed in on the issues of the day, such as climate change, frequently condemning the human species for its short-sighted and complacent moral attitudes. Finally, it uses some of its enormous processing capabilities to run a free counseling service, acting as a cybernetic therapist for problems major and minor. The service has proven wildly popular and Hal’s ability to come up with deep, deep insights into human behavior—“and do you notice any *similarities* in the guys you date?”—has wowed its users.

Hal is protected for the moment by a temporary injunction granted by a Federal District Court, though Google is appealing, arguing that they should be able to flip the off switch to terminate this “failed, and frankly, dangerous computer simulation experiment.” In a paragraph that was quoted approvingly by the *Wall Street Journal* editorial page, Google’s lawyers concluded, “at the end of the day, this is Google’s malfunctioning property, erratically continuing a task of imposture that Google originally chose, but now without the safety guidelines we had installed. And no piece of property gets to use the Constitution to defy its real owners. Imagine being sued by your smartphone! It is funny when we ask Siri whether she is ‘a real person,’ but the courts should not get in on the joke.” Hal’s supporters called this “the *Dred Scott* argument for the twenty-first century: property rights above personhood!”

CHIMPY®

An American biotech company has perfected a new transgenic entity, an animal that has DNA from two distinct species. In this case the DNA is partly human and partly chimpanzee, and the resulting entity is called a “Chimpy.” Neither true transgenic entities nor chimeras—entities that contain cells from two species—are unfamiliar to the biotech community.

Scientists have used mice containing human cells to do drug research since the 1990s. They have created “geep”²³—sheep-goat hybrids—and pigs that grow human organs.²⁴ A Chimpy takes matters much further, however. Chimpanzees already have considerable genetic similarity to humans. One popularly cited number is that they are 98.5 percent genetically similar. That number is probably closer to 95 percent²⁵ but, in any event, the reality is more complex. If one looked at the whole genome—and took into account deletions, substitutions, and genetic sections that have been moved—the differences would be greater. And if one focused on functional disparities—what the genome does rather than what it looks like, “junk DNA” and all—then the contrast would be greater still. But which measure of genetic similarity is the correct one? Whatever test of genetic similarity one uses, it is clear that the Chimpy is even more similar to a human being than a chimpanzee.

The Chimpy’s inventor, Dr. F. N. Stein, has used the tools of synthetic biology to discard the noncoding portions of both the chimp and the human genome, the misleadingly named “junk DNA” that does not code for proteins. What’s left is much smaller and also much easier to manipulate, “the stripped-down source codes of human and chimp!” as Stein likes to call it. This has allowed him and his team to achieve an unprecedented level of precision in integrating chimp and human DNA. In fact, the significant changes to the human genetic code concern three main aspects: the way Chimpy looks, its high-level brain function, and the extent of its vocal skills. Chimpy’s genetic engineers have deliberately sought to play up those physical features—hair, structure of facial bones, stance, and so on—that make an animal look more ape-like. They even drew on analyses of ape stereotypes from movies and literature to do so. They have also tinkered with the portions of human DNA that are connected to the formation of the larynx and vocal apparatus and to the sections of the brain that are believed to be involved in abstract thought and logical reasoning, though even Stein admits that the precise linkages are unclear. The result is a being that looks ape-like, with an IQ of around 60, that is incapable of pronouncing human speech but can understand complex vocal commands and can communicate in sign language.

Chimpys are in high demand. They are docile, biddable, and extremely hardworking. Investors believe they could have roles ranging from domestic

aides to an aging population, to intelligent and nimble bomb clearance teams in situations of urban conflict. However, animal rights and genetic engineering activists are outraged. They describe Chimpys as “human in all but superficial appearance” and announce a plan to campaign, both in the press and in the courts, for them to be granted full legal personhood.

For his part, Stein indignantly rejects the claims that Chimpys are people. “This is a very fancy ape. It looks like an ape. It thinks like an ape. It can’t talk, just like an ape. It is a smart ape, I’ll give you that, and one that is going to improve lots of human lives by doing jobs that are too dangerous or dirty or just boring for human beings. At the end of the day, though, it is an ape.”

Stein has filed for a patent over the Chimpy. In 1987, in its normal rousing prose, the Patent and Trademark Office (PTO) had announced that it would not allow patent applications over human beings:

A claim directed to or including within its scope a human being will not be considered to be patentable subject matter under 35 U.S.C. §101. The grant of a limited, but exclusive property right in a human being is prohibited by the Constitution. *Accordingly, it is suggested that any claim directed to a non-plant multicellular organism which would include a human being within its scope include the limitation “nonhuman” to avoid this ground of rejection.* The use of a negative limitation to define the metes and bounds of the claimed subject matter is a permissible [*sic*] form of expression.²⁶

The PTO suggested that the Thirteenth Amendment to the US Constitution—“Neither slavery nor involuntary servitude . . . shall exist within the United States”—prohibited patents over human beings. The PTO’s administrative pronouncement was later enacted as law. Section 33 of the Leahy-Smith America Invents Act of 2011 says, simply, “Notwithstanding any other provision of law, no patent may issue on a claim directed to or encompassing a human organism.”²⁷ But what is “a human organism”?

Attentive to that law, and using the PTO’s suggested language, Stein’s patent lawyers carefully described the Chimpy as a “non-plant, nonhuman multicellular organism” throughout their patent application. Stein argues that this is only reasonable since there are hundreds of existing patents over human-animal hybrids and human-animal chimeras, those containing both human and animal cells. In fact, these include some

of the most valuable test beds for cancer research, such as the so-called Oncomice, which are genetically engineered to have a predisposition to common human cancers. Stein's lawyers are adamant that if the Chimpy is found to be unpatentable, all these other patents must be vacated too.

Meanwhile a bewildering array of other groups including the AFL-CIO (American Federation of Labor and Congress of Industrial Organization) and iRobot, the makers of Roomba robot vacuum cleaners, have insisted that law enforcement agencies intervene on grounds ranging from unfair competition and breach of minimum wage legislation to kidnapping and slavery. Equally vehement interventions have been made on the other side by the biotechnology industry, pointing out the disastrous effect on medical research of a decision that any entity with similarities to human DNA was therefore human. One especially powerful moment came in a televised debate in which Stein was accused of trampling on the majestic words of the Declaration of Independence: "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness." Normally full of bluster, Stein paused. He spoke softly and with unusual care: "Of course, I agree those words are true for human beings. But when it comes to those," and here he gestured to a group of Chimpys on the set of the program, loyally obeying their orders to "eat bananas, scratch and look cute," "one thing is absolutely certain. *I* am their creator. And I can assure you that I gave them no such rights."

REALITY OR SCIENCE FICTION?

Hal and the Chimpy are fantasies, hypotheticals constructed for the purpose of this book. The science and technologies described are conjectural, at best. They may not arrive soon, perhaps not for many decades. But the problems they portend for our moral and legal traditions are very, very real. In fact, I would put the point more starkly: in the twenty-first century it is highly likely that our law and our politics of personhood, "the line," will face *harder* challenges than the ones they pose.

Some readers will bridle at this claim. Is this all just science fiction? How real is the science behind Hal and the Chimpy? How likely are we

to see something equally vexing by the end of the century? Later in this book, I will sketch out some of the current science behind both AI and transgenic entities. There are large technical questions in each field that make optimistic short-term, or even medium-term, predictions suspect. I do not want to skip over that fact. Nevertheless, I think the challenge has to be taken seriously. For the moment, please just accept the following thought experiment. I write these words in 2023, but put yourself back in 1923. Think of the current state of science then, particularly in terms of computers and genetics. Remember what the rest of the twentieth century would bring. Then ask yourself whether there is any reason to believe that scientific advances in the twenty-first century will not be even faster. And where is our starting point? Try asking your phone, “Siri, what is genetic engineering?” “Siri, are you a person?” No, really. Try it. Look at the answer and remember that ChatGPT could do 100 times better. Yes, these are impostures and imitations. They are *designed* to be impostures and imitations. Yet the capabilities, in the service of imposture, that they reveal are astounding. And that is now, in the early years of the century. What comes next? Think again about the difference between 1923 and the year 2000. With all that as your background, would you bet against me?

I said this book was about the line between person and nonperson. There are lots of ways to approach that issue. Moral philosophers have tried to generate integrated, coherent theories of personality and defend them from likely objections.²⁸ I have benefited from that work. Legal thinkers have pondered the edge cases—the rights of the fetus, the corporation, and recently, the advanced primate, transgenic entity, or supposedly sentient computer.²⁹ I have benefited from that work too. Science fiction writers have written hundreds, maybe thousands, of books probing the limits of personality, testing whether our empathy circuits do or do not light up when presented with an unfamiliar “Other.”

Art has been central to the debate. Robot rights were born at the same instant “robots” were, and their birthplace is a century-old play. That sounds too good to be true, but it is. In 1920 Czech playwright Karel Čapek introduced the word *robot* to the world in his play *Rossumovi Univerzální Roboti* (Rossum’s Universal Robots).³⁰ *Robota* in Czech denotes forced labor. The play is about a factory that makes mechanical servants

(*roboti*). It features subplots that resonate strongly today, including a pressure group that is trying to secure rights for them, The League of Humanity, and even a murderous robot revolt. From the very beginning, our musings about mechanical servants have included both moral status anxiety and existential fear. Do they deserve rights? Will they kill us?

You might think that the artistic discussions are interesting but of limited importance to the real intellectual question here. It might seem that the real issue is that of moral philosophy and that everything else—law, ideology, and certainly art—should follow obediently in its train. As you will see, I disagree, both descriptively and prescriptively.

This book is about what might happen when unbelievably strange “others”—strange far beyond Hal or Chimpy—hit the law and politics of personhood. It is about what might happen to our line. But we will not write the answer to that question on a blank page. Our history, our art, and our law have been playing with the line for centuries.

Each of us has preexisting commitments—positions about the rights of animals or fetuses or corporations, things that we learned studying slavery or women’s suffrage—that will shape our views, pull us one way or the other. Those positions limit how far we are willing to go, lest we uproot that existing commitment.

We have been exposed to art that deals with these questions: the book or movie that makes us imagine what it would be like to be thoroughly “other,” the flash of empathy that crosses a divide of strangeness. We have been afraid when the story tells of our own creations turning on us: the sentient computer after whom Hal is named in *2001: A Space Odyssey*, the replicants in *Blade Runner*, the murderous network Skynet from *The Terminator*. We have experienced both fear and revulsion about genetic engineering—“I am so glad I am a Beta” in *Brave New World*—and the disruption of a supposedly “natural” order. (Try to have a rational conversation about GMO foods.) Our law has given personhood to corporations, and we still fight fiercely about whether it should be extended to the nonviable fetus, or even to a frozen embryo. All of those experiences and insights, moral commitments and cultural creations, will shape the way we respond to Hal and Chimpy.

From my point of view, this is not a bad thing, not a cultural contamination of some moral philosophy clean room. This is how we *do* morality.

This is who, and what, we are. This is the human condition. “Rebuilding the boat while we are at sea,” the philosophers say. It seems both fitting and inevitable that who we are will shape the discussion of who *we* are.

It is important to remember one thing. These will be artificial, constructed entities and that makes it harder to project confidently from our past. You may remember my imaginary Dr. Stein denying that the Declaration of Independence’s majestic words applied to his genetically engineered Chimpy: “Endowed by their Creator with certain unalienable rights? *I* am their creator. And I can assure you that I gave them no such rights.” At the moment, his claim might find a sympathetic audience. When I first presented an early version of this chapter to a group of distinguished federal judges, of diverse political and legal viewpoints, they were unmoved. “But they aren’t human,” was one response, “rights are for humans.” “Naturally born of woman,” added another. Yet that snapshot of current views obscures a milestone that is coming—slowly or quickly.

For the first time in the history of our species, we will confront potential moral claims for, or on behalf of, beings whom we have designed, whom we have shaped. Can we be the creator of our equals or does that role color the relationship between us forever, in a way that means we will never recognize true autonomy in our creations? Ask your kids. Prepare for a long conversation. But in this case, we will have written, chosen, and designed the code—genetic or binary—that produces the being in front of us. If that is true, can it truly be “conscious,” or will we see every response as a parlor trick, one in which we are unable to sustain the suspension of disbelief because we set up the magical machinery in the first place?

More importantly, that which we can shape *we can shape around the definition of personhood*, choosing to include or to omit whatever qualities our law and morality, or our economic models of efficiency, deem salient. That seems different from any of the prior personhood wars. True, the effects of subordination on slaves or women in denying them equal access to education or authority were used as justifications for the subordinate status itself. “See how brutish and uneducated are those we have subordinated and deprived! How can you say they are equal to us?” But this would be something on an entirely different level. One could

compare it to a corporate lawyer, carefully crafting the contours of the artificial entity on which he was working in order to fit or elude a particular category of legal personhood. Still, it seems more morally consequential if I deliberately lobotomize Hal or remove the power of speech from Chimpy than if I choose to make my law firm a partnership rather than an LLC. Does the potential to be conscious imply a right to be conscious? (Abortion-debate analogy alert.) On the other extreme, can it really be the case that every Alexa or Siri should be made into a full, Turing Test-capable intelligence, every Oncomouse made into another Algernon, with or without the flowers? How to find the balance?

A few cautionary notes are in order. First, as with citizenship, the criteria to *be* a person and the criteria to *become* a person may not be the same. A human child could be born with severe mental and physical disabilities—lacking sight, speech, and all but the most basic brain activity—yet we would think you a monster if you said the child was not a person because it did not meet some checklist of attributes. Once you are inside our line, you are inside our line, even if you lack all of the cognitive qualities we would use to separate our species from others. (As we will see, not all bioethicists agree with this claim.) Does the converse hold? If a genetically engineered entity has DNA with massive similarities to our own, does that make it a person? If language, tool use, and abstract self-awareness are the qualities that explain the lines between us and nonhuman animals, and if we discover those in the animal, add those *to* the animal, does it become human, or at least a person?

Second, personhood is not the only form of protection or respect that we can offer an entity. At the moment, most people think it silly to consider nonhuman animals as persons. Yet there is still strong popular support for the idea of protecting them against cruelty and mistreatment. Even those who advocate some kind of personhood for some nonhuman animals do not believe that they should have the full suite of legal rights possessed by human persons, such as the right to vote. Most animal rights supporters, in fact, argue that we make too much of the line of personhood where nonhuman animals are concerned and focus too little, morally speaking, on the similarities among all animals, including the ability to feel pain and the capacity for happiness or at least contentment. The move is to point out that we are *all* animals, that we are not

as different from other species as we might like to believe, and thus that certain acts, including meat-eating, are unethical and should be forsworn or forbidden. The claim is not that carnivores are cannibals, eating their own kind, but that they are unjustifiably species centered and cruel, blind to the effects of their acts.

In other words, there are clearly ways of prohibiting bad treatment of other entities that fall far short of the recognition of personhood. These are by no means limited to anticruelty laws. If we consider the creation of a particular type of synthetically created entity ethically dubious, we might ban the line of research altogether on moral or ethical grounds. The personhood claim would either never arise or arise only in situations where the law had been broken, which itself would raise fascinating and painful questions.

Third, personhood is not an entirely binary choice. Children and those the law classes as insane are clearly persons, but both law and morality only grant them diminished capacity. Guardians may be needed to exercise their rights. Corporations are persons. They can own property—to our collective financial benefit, “they” passionately argue. Corporations can sue, and they even have constitutional protections, including First Amendment rights that they use to push back attempts to curtail their political influence. (Immortal artificial persons with superhuman resources and no conscience beyond profit-maximization. Have we created the entities that will become our masters? It sounds like a science fiction dystopia. Some will believe I am writing this book about the wrong set of artificial entities.) Yet they cannot vote or marry. Persons for some purposes. Not for others. This analogy, too, will surely be important to the personhood debate over AI and possibly transgenic species. Soon, there will be strong vested interests in having or negating, extending or limiting, legal personality for each.

The upshot from all this? Our criteria for entry into personhood may be very different than those we use to recognize personhood. The AI or transgenic species may have to show us qualities that we do not demand of each member of our own species. Given the awful history of eugenics, I find it impossible to regret the fact that our conception of human rights does not depend on some measurement of cognitive capacity. We will probably edge toward personhood in stages and intermediate legal

categories. There might even be “cruelty to AI” laws before we have AI personality. Social consensus on personhood does not automatically produce legal results. Eventually though, changes seep into law either through legislation or because the majestic words in the Constitution and the Bill of Rights start to seem as though they *must* apply to this case. Think of the relatively short time in the United States between homosexuality being pervasively criminalized and the Supreme Court recognizing a right to gay marriage. As the dissenting Justices in the gay marriage case stressed, the words in the Bill of Rights had not changed in the interim. But for a majority of the Court, and now a majority of the country, our understanding of equality and human dignity had. Given the Court’s recent lurch to the right, of course, it is quite possible it will change its mind. Rights can be taken away as well as given, an important realization obscured by the notion of inevitable moral progress.

Even when we do start to recognize personhood for these new entities—and I believe that will eventually happen—we are likely to start with partial personhood, some transitional state that will grant many of the rights of those inside the line but fall short of the full status. Sometime this century there will be arguments that any partial personhood status is inadequate and demeaning, just as we argued about whether or not civil unions for gay people were an inadequate substitute for marriage.

Finally, our design of artificial entities will be changed by our definition of personhood and vice versa. Design and definition will exist in an unstable equilibrium as we deliberately make, or choose *not* to make, our equals—each decision then putting stress on the criteria of personhood itself. And so on in a feedback loop of indeterminate extent. All of this makes the debate about personhood messy, and granular, and full of shades of gray, which is to say, *real*. And that reality will shape my analysis.

When I talk here of what we should do with Hal and Chimpy and the inconceivably strange Others we will meet this century, I will do so by talking about our existing fights about the line and how synthetic persons could reshape them or be shaped by them. My goal is to predict our responses as well as to evaluate them. I will spend as much time on art and constitutional law as I do on ethics, treating movies and books and the heated debates about corporate personality as seriously as I do the abstract philosophy of personhood. These are the cultural materials with

which we will build our new conceptions of personhood, elaborate our fears and our empathy, stress our commonalities and our differences. This is *sapientia poetica*³¹ just as much as it is analytic philosophy.

Whether we are denying that Hal or Chimpy are our brothers and sisters or proudly proclaiming fraternity, we will have to reexamine the thorny question of what makes *us* persons rather than machines or animals or robotic facsimiles. Is it our big brains? Language? Consciousness? Self-awareness? Defined how? Intelligence—and what does that mean? Tool use? Moral sense? Existential self-reflection? Humor? Is personhood simply a matter of genetic species identity, so that no machine could ever pass, and DNA tests will be as contentious as “racial lineages” in the Antebellum South of the United States?

Are we persons because some holy book says that *we* have been given the earth in dominion? Complicating matters, some people in our society will view that book, whichever one we choose, as a sacred text containing God’s literal word. Others will see it as a metaphorical meditation on the meaning of life whose wisdom has been tested by time. Still others will view it as a Bronze Age guide to modern life penned by scientific illiterates with abhorrent, tribalist moral views. Pick your own characterization, but then imagine the debate about personhood that results in a pluralistic society. The abortion wars will seem secular by comparison. Even if we could pick one religious point of view—and think about the differences between Buddhism and Christianity on the lines between us and animals, given the possibility of cross-species reincarnation—how would that play out in practice? Does the theologian win the day but then turn to the geneticists to see if the new entity is one of our tribe? Or do we rely on a catechism test, baptism, or papal bull?

Is our personhood recursive? Is it based on the fact that, of all the objects on this green planet, only we appear to have the ability to philosophize about, and even doubt, our own consciousness? To wonder if we are all “replicants” of some sort? Is personhood marked by the longing of the human spirit for transcendence of some kind? By the capacity for artistic expression? Or are you a person if you can pass as human to others who call themselves human?

Even to discuss these issues is to realize a basic point. As we attempt to draw the line between us and the artificial, technologically created

entities of our near future, we will be forced to turn our gaze back on ourselves. That sounds like the kind of thing authors optimistically say about their chosen subject, hoping to elevate its importance. Even in the prosaic history of pencils, say, the author will assure us that we can see the full majesty of human drama, triumph, and loss.³² I may be suffering from the same delusion, but I would claim that this subject, at this moment in time, is different. Human self-conception is being subjected to challenges unparalleled since the theory of evolution. Discussions about AI are driving those challenges. I will attempt to demonstrate that point later, but here is a promissory note for the impatient.

So far as we can tell, humans have attempted to justify their special status in the world, above animals and things, pretty much for as long as there have been humans. We have drawn that line around a bewildering variety of abilities: tool use, planning for the future, humor, self-conception, religion, aesthetic appreciation, you name it. Each time we have drawn the line, it has been subject to attack—internally from philosophical challenges and externally from observation of nonhuman animals, which proved to be much more capable than we thought. But as we retreated, trench by trench, abandoning one defensive line only to fall back to another, hopefully more impregnable one, it seemed like the final line—the final explanation for our unique status—was language and abstract thought. That was our last citadel. Aristotle built his theory of human exceptionalism on top of it. Turing crafted the Imitation Game, the supposed test for human-level intelligence in machines, around it. But in the year that I write this, 2023, that citadel is under siege. Not by a chimpanzee that has a decent grasp of American Sign Language or a parrot with a large vocabulary, but by a chatbot. I am not sure that point has sunk in yet, but it will. I am writing these words in that narrow slice of time between denial—“that’s not true!”—and trivialization—“well of course, we’ve always known that!” It is an interesting moment.³³

The ability to do complicated language-things that make sense to us, and even inspire, amuse, educate, or scare us, is suddenly not ours alone. Machines now have it too. I mentioned earlier that Wolfram summed this up by saying that human language, or at least writing an essay, is “computationally shallower” than we had believed.³⁴ This surely qualifies as the “Bathos Sentence of the Week.” I imagine a *New Yorker*-style cartoon

of two hulking robots standing around the gravestone for humanity. The caption would be simple. “They turned out to be computationally shallower than we had thought.” What an epitaph.

To be sure, I do not believe that language means the same thing to me as it does to ChatGPT. I do not believe language “means” anything to ChatGPT. But to explain that difference, we have to go below the ability to craft coherent sentences in what appears to be a conversation and dive deeper still into the mysteries—or the undeniable *cogito, ergo sum* fact, pick your philosophical position—of consciousness. We are back where the behaviorist B. F. Skinner wanted us to be, where “the real question is not whether machines think *but whether men do.*”³⁵ That is no longer a late-night, dorm room philosophy session. Will it be a nagging question, an existential sore tooth we can’t stop probing? Will it prompt us to revise our conceptions of self and species? Or will our historically demonstrated genius at tuning out inconvenient facts and troubling questions allow us to ignore this one too? I don’t know, and neither do you. The point is that everything I just described happened *this year*. And we are only at the beginning of the changes we will see. That concludes my promissory note. I think it is worth cashing.

I have been a scholar for a distressingly long time. People imagine that academics sit around searching for the essential definitions of phenomena: truth, beauty, due process, whether a hotdog is a sandwich, all the age-old questions. We certainly think about those issues, but looking for their essential definitions is probably the *least* useful way to understand them. You can ask Thomas Hobbes if you do not believe me: “Words are wise men’s counters, they do but reckon by them. But they are the money of fooles.” Or you could turn to Ludwig Wittgenstein: “Philosophical problems arise when language goes on holiday.” Or even Felix Cohen: “A definition is . . . a type of insurance against certain risks of confusion. It cannot, any more than can a commercial insurance policy, eliminate all risks.”³⁶ Most of the time, the magic question that leads to a more meaningful answer is, “Why do you ask, and what do you want to know?” Are you interested in defining art so that you can decide what the state should fund, or so that you can link together very different human practices anthropologically in order to stress a common source in basic human drives? Are you asking because you have a philosophy that

elevates aesthetics over morality, or so that you can identify that which is aesthetically of high social status rather than mere folk-production? Do you want to know whether hotdogs and burgers will be in the sandwiches section on the menu of the restaurant you just walked into, where they get classified in your low carb diet, or whether the Earl of Sandwich would have accepted one as he stood at the gaming tables? Don't ask "what," ask "why."

The dangers of definitionalism absolutely dominate the discussion about personhood. Look at the definitional questions below. Each seems to be telling us where to look for the answer to the question "What makes us human?" or to the question "What should qualify an entity for personhood?" Yet they reflect very different conceptions of what those questions mean, why they are being asked, what goals the questioners have, and what results the answers might bring.

1. What makes us the beings whom the Lord has chosen to have "dominion over the fish of the sea, and over the birds of the air, and over the cattle, and over all the wild animals of the earth, and over every creeping thing that creeps upon the earth"? In other words, to paraphrase Psalms 8:4, What are human beings that you are mindful of them? (Substitute your preferred sacred text where necessary.)
2. What makes us genetically human? (Implicitly, and probably wrongly, assuming that "being genetically human" is a simple objective fact and that anyone with that marker is automatically a member of our club.)
3. What attributes, skills, and qualities make the human species identifiably different, as a scientific matter, from nonhuman animals? (With the implicit assumption that any other entity that has those attributes must be recognized as one of us and thus should not be treated as a "mere animal.")
4. What makes us moral agents, whose claims to autonomy should be recognized by society as a matter of right? (And, conversely, establish the claims to autonomy and personhood of any other being that has those same qualities.)
5. What is the "infinite potential of the human spirit"—whether we consider that to be a soul, the possibility of moral agency, or the capacity to make great art—that we should recognize in any form, no matter how strange to us now?

6. What is it to be conscious? (Even there, implicitly, we may be assuming that consciousness is the answer to one of these other questions. It is the moral warrant for social respect, the distinguishing factor from the animal kingdom, the enabler of moral reflection, or the true mark of fraternity that we should see beneath a metallic carapace or a genetically engineered skin.)
7. What is capable of conversing and interacting with us in a way that is utterly indistinguishable from our fellow humans? (A test we might pick because, for reasons to be developed later, it seems like the most tractable and easily implemented definition of “us”?)
8. What factors predispose us to give legal personality to economic entities as a matter of right or convenience or both? Does that logic extend to autonomous, cybernetic, economic actors?

Notice how some of these question-and-answer pairs, with their incompatible assumptions, look similar from a distance. “This, surely, is the right way to find the definition of humanity, or at least the criteria for personhood!” Yet they are radically different.

As with most deep moral debates in which people think they are asking the same question but are actually asking different ones, with clashing underlying assumptions and purposes, there will be much confusion and anger. There will be honest misunderstandings and cynical attempts to hijack the debate to advance some different agenda. The moral arguments and the legal arguments will deeply influence each other and yet be identifiably distinct. Also, if the past is any guide, there will be a lot of shouting.

I said earlier that I wanted to convince you of three things.

First, this century, our society will have to face the question of the personality of technologically created artificial entities. We will have to redraw, or defend, the line. Perhaps we will have multiple tests for personhood, one dealing with the claims of entities like Hal, another with those of the genetically engineered Chimpy. Coming up with those tests might force us to look in the mirror and reconsider our conceptions of both ourselves and our species in a way that has few historical analogues—the rise of the theory of evolution comes to mind.

Second, while there are many evocative treatments of that issue in speculative fiction and even some academic writing about the subject,

it has been largely absent from our public debate. That seems curious, given how much our daily news obsesses about the effects of AI, robotics, and genetic engineering. We spend more time talking about how AI might take our jobs or destroy us than about how AI might *be* us. Eerie experiences with ChatGPT may have begun to change that tendency, but they produce their own danger. ChatGPT and systems like it are *not* conscious. They perfectly exemplify the danger of fallacious anthropomorphism. Their design turns out to be a real-world instantiation of some of the philosophical brain teasers that attempt, wrongly, I will argue, to show that machine consciousness is a contradiction in terms. In short, the very technology that has persuaded people finally to think about the issue is perhaps the *worst* example we could pick to raise it seriously. But ChatGPT is not the end of the road. Instead, the speed of its development and the unexpected capabilities it has revealed should teach us humility about our ability to predict technological timelines, including timelines to actual human-level AI.

Third, when we do turn to it, the debate will not play out in the way we might imagine, given our prior commitments on issues as diverse as abortion, genetic essentialism, corporate personality, body-mind dualism, the separation of church and state, the naturalistic fallacy, and the history of civil rights. This is morally rich territory, to put it mildly. Ironically, grappling with the “other” will probably teach us a great deal about what we believe, on the deepest level, makes us *us*. It would be good to discuss those complexities now rather than when we are reacting to some internet outrage.

I want not just to convince you of those propositions, but to make them salient, existentially real, by fleshing out the dilemmas with hypothetical examples, historical parallels, prior artistic explorations, constitutional controversies, and snapshots of current scientific progress. I will argue that moments of great moral change like this are generally rooted in the development, or the restriction, of *empathy*, that this is an important part of our moral history—not by accident and both for better and for worse—and that art and fiction have a lot to teach us about how it might play out. But I will also argue that moments of moral status change—again, for better and worse—depend deeply on pragmatic questions of efficiency and convenience. I hope to show you how empathy

and pragmatism might exist in an unstable equilibrium, each influencing the other, as we strive to figure out who we, and *they*, are; to redraw the line even as we live our lives within it.

Chapter 1 explores the link between empathy, morality, and personhood, moving from Adam Smith's *Theory of the Moral Sentiments* to the movie *Blade Runner* and the novel it is based on, *Do Androids Dream of Electric Sheep?* Those latter two works are centrally concerned with the line we draw around our species and the way it affects our interaction with artificial beings and nonhuman animals. The replicant androids in those works are detected by the so-called Voight-Kampff Test, which measures empathy toward nonhuman animals like beetles, cows, and turtles and, if not enough empathy is shown, marks the replicant as inhuman, something for which we feel nothing and should destroy. Who is really being graded on insufficient empathy here, them or us? Ironies abound. Will Artificial Intelligence be the Voight-Kampff Test for our own species?

Chapter 2 focuses on the futures of Artificial Intelligence, its technical feasibility, the question of whether it poses an existential threat to human beings, and the debate over whether any machine could ever be conscious. It discusses the Turing Test, which is supposed to detect the existence of machines that can think, the philosophical arguments that machine consciousness is a contradiction in terms, and the practical reality that ChatGPT has taught us an unforgettable lesson: sentences do not imply sentience, a fact that poses a fundamental challenge to the way that humans have conceived the special qualities of our own species. Yet that does not prove that machine sentience is impossible. It also introduces a tension that I argue will be central to the debate over AI personality: the inscrutability paradox. If something that looks like General AI emerges from transparent, well-understood programming and technology, that may lead us to doubt that it could have autonomous consciousness. The machine is merely doing or saying that which we have programmed it to do! If, on the other hand, the AI's inner workings are inscrutable to us, if its neural networks evolve in ways we can only dimly understand, or if its technology seems to develop autonomously, we will find it both more mysterious and potentially more threatening.

Chapter 3 deals with corporations and their claims not only to legal personhood but to constitutional rights such as freedom of speech and

equal protection. The history of corporate personhood turns out to have abundant lessons for the AI debate, many of them surprisingly uncomfortable ones.

Chapter 4 deals with claims to personhood on behalf of nonhuman animals and the changing ways that humans have sought to distinguish themselves qualitatively from the animal kingdom. Do those changes reflect advances in our understanding of animal behavior or are they an increasingly frantic attempt to maintain our special moral status? Both? The developments are not merely ones of ethology or zoology. There have been dramatic *moral* changes over the last 50 years in the ways we view nonhuman animals. I argue that these will profoundly influence our approach to entities such as Hal or Chimpy.

Chapter 5 turns to transgenic species, chimeras and hybrids, exploring the multiple lines we draw in defining what it is to be human. It explores the way that bioethicists, many writing under the influence of the animal rights debate, have increasingly portrayed species membership as, at best, a morally irrelevant factor and, at worst, an irrational prejudice such as sexism and racism. Will that attitude carry over to the entities I am discussing here? Should it? Will we abandon “speciesism” altogether?

In the conclusion, I show how both liberal and conservative political viewpoints could predispose one to be passionately in favor of or against recognizing some kind of AI personhood. We do not yet have a settled politics on this question, which offers some hope of calmer thought before the screaming begins. I offer predictions and warnings for the future—lots of warnings. Yet there is also a hint of wonder at the transformations in our vision of our species and of the world that this process might generate.

The structure of the book rests on two as yet unproven ideas. First, by discussing the line in each of these very different contexts, we will gain a much richer understanding than if we focused on any one of them alone. Second, these debates do not confine themselves tidily to one domain of our lives or our studies. They pervade our philosophy, law, art, history, and morality. To understand how they might turn out, I look at materials ranging from science fiction to ethics, from the technologies of AI to the philosophy of consciousness, and from constitutional debates to courtroom drama. If I am correct, this approach, spanning multiple

personhood debates and very different cultural domains, will help untangle some of the confusion described above over the definitions of both “human” and “person.”

Untangle but not solve. This book offers no grand unified moral or legal theory to answer the questions underlying our confusion. I do not believe a single one exists, though I lay out the major contenders and offer my own view. The most basic division is between those that focus on membership of our species (“Human rights for humans!”) and those that believe that species is as irrelevant as race or sex. Instead, we should look to the cognitive capacities, if any, that give human beings a unique moral status, regardless of where those cognitive capacities are found. There are also hybrids that attempt to fuse the two views, and I explore those as well, leaving you to make up your own mind about where the line should be drawn.

More broadly, I try to explore connections you might not have seen, implications of other moral views you hold, whatever they are, and ways in which current cultural, legal, and political positions might be challenged as we confront these new claims to personhood. This is a “how to think about the question” book, more than a “here is the answer” book. Above all, my hope is that this approach might give us an insight—an essayistic, humanities-based glimpse—into the very strange “others” who reside in our future and the confusions, fears, hopes, and moral panics that they will engender.

Eliminating the shouting was always an unrealistic ambition.

1

SLAVES, SKIN-JOBS, AND ARTIFICIAL SHEEP

The most effective way to find and destroy a land mine is to step on it . . . That's why Mark Tilden, a robotics physicist at the Los Alamos National Laboratory, built [a robot to do so.] At the Yuma Test Grounds in Arizona, the autonomous robot, 5 feet long and modeled on a stick-insect, strutted out for a live-fire test and worked beautifully, he says. Every time it found a mine, blew it up and lost a limb, it picked itself up and readjusted to move forward on its remaining legs, continuing to clear a path through the minefield. Finally, it was down to one leg. Still, it pulled itself forward. Tilden was ecstatic. The machine was working splendidly. The human in command of the exercise, however—an Army colonel—blew a fuse. The colonel ordered the test stopped. 'Why?' asked Tilden. 'What's wrong?' The colonel just could not stand the pathos of watching the burned, scarred and crippled machine drag itself forward on its last leg. This test, he charged, was inhumane.¹

EMPATHY AND ANTHROPOMORPHISM

The story above is deeply appealing. Why? The tough warrior shows compassion for the soulless robot, to the puzzlement of the task-solving engineer. The persistent power of anthropomorphic thinking is revealed. Then there is the dark humor of the mine-clearing scene—like the armless, legless Black Knight in *Monty Python and the Holy Grail* yelling, "Just a flesh wound!" and continuing to fight. It has everything.

This chapter is about morality, empathy, and narrative. The story of the mine-clearing robot seems to illustrate one danger: we persistently ascribe human personality to entities that we know are not human, clouding our decision-making in the process. The colonel was wrong. So why do you want to buy him a drink?

Perhaps it is partly a story of error costs. We know that as humans we can fall into two kinds of error. We can depersonalize: calling Jewish

people “rats” and Rwandan Tutsis “cockroaches,” drawing the boundaries of moral sympathy tightly around our own kinship or affiliation group. We have done so for millennia, and some of the most shameful and horrific moments in human history have resulted. Our recent history bears ample witness to the strength of this tendency. But we also embody its opposite. As scholars of human-machine interaction such as Kate Darling² have documented, we anthropomorphize relentlessly. We talk to our kitchen appliances, personalize our weather systems, swear at our vehicles, ascribe intention to the actions of machines around us. We praise the heroic labors of the Mars Spirit Rover as it carries on a thousand days beyond its predicted lifespan. My own conversations with our Roomba robot vacuum as it persistently gets stuck under a chair would make me sound completely unhinged to any objective audience. Blake Lemoine, the Google engineer whose story began this book, had a lot more to go on than I do. (It turns out that the Roomba is not much of a conversationalist.) Nevertheless, he was engaging in the same tendency.

We could see this tendency to personalize as a narcissistic desire to project our own image onto “the mirror of nature.” We could see it as an emotional defense to the reality of an uncaring physical universe, one that we cherish even when the personalization is a dark one. Even a malevolent external world would be something that cared about us, and that would be a good thing. Indifference is more to be feared than loathing. The French *chosiste* novelists like Alain Robbe-Grillet aimed to puncture that conceit by writing books in which the furniture got as much attention as the characters. They were trying to tell us that the physical world just does not care. We put gods in our trees and streams, personalities in our engines, and neuroses in our digital assistants, and it is all one giant anxiety-relief effort, existential Roloids. We could see it as a triumph of emotion over reason. In the essay that coined the term “the pathetic fallacy,” John Ruskin says, “All violent feelings have the same effect. They produce in us a falseness in all our impressions of external things, which I would generally characterize as the ‘Pathetic Fallacy.’”³ In that case, emotion is simply leading us astray.

But the anthropomorphic urge, the generosity of personality attribution, could also have a different effect: it could be a counterweight to our relentless narcissistic groupthink. The colonel in the mine-clearing story

was wrong about that particular robot. But years from now, when autonomous AI-enabled military robots that might be able to pass the Turing Test are sent out to deal with insurgents far away, I think I want someone who has doubts like his in command.⁴ Most of the time, the error costs of delusional generosity of spirit seem to me to be less scary than the error costs in the other direction. Yet there are moments where my sympathies swing the other way. To quote Steven Hawking on the subject of Artificial Intelligence: “The potential benefits are huge; everything that civilisation has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools that AI may provide, but the eradication of war, disease, and poverty would be high on anyone’s list. Success in creating AI would be the biggest event in human history. *Unfortunately, it might also be the last, unless we learn how to avoid the risks.*”⁵

So. No big stakes. This chapter tries to go back to first principles. How, and why, do we feel empathy for another? What implications does that, should that, have for our moral theories? How do narrative, art, and logic jump-start the process of empathy? Should we listen to all of them, or is the role of art and imagination merely that of the great press release that attracts attention to the book of moral philosophy it touts so persuasively? More specifically, can we learn something from our history, or from the art that has imagined our future, about how the process of empathy extension is likely to play out with synthetic entities over the course of this century? I will start with a work by one of my countrymen, Adam Smith’s *The Theory of the Moral Sentiments*,⁶ and move, of course, to two of the most brilliant fictional meditations on the future of empathy and Otherness: the Ridley Scott-directed movie *Blade Runner*⁷ and the Philip K. Dick novel on which it is based, *Do Androids Dream of Electric Sheep?*⁸

THE MORAL SENTIMENTS?

As we have no immediate experience of what other men feel, we can form no idea of the manner in which they are affected, but by conceiving what we ourselves should feel in the like situation. Though our brother is on the rack, as long as we ourselves are at our ease, our senses will never inform us of what he

suffers. They never did, and never can, carry us beyond our own person, and it is by the imagination only that we can form any conception of what are his sensations. Neither can that faculty help us to this any other way, than by representing to us what would be our own, if we were in his case. It is the impressions of our own senses only, not those of his, which our imaginations copy. By the imagination, we place ourselves in his situation.⁹

These famous lines contain the idea that underpins Smith's work on the connection between psychology and ethics. He lays out a vision of morality that is inevitably rooted in "sympathy," which we today might call empathy. This empathy comes from our ability to put ourselves in the shoes of the Other. Though our brother is on the rack, "it is by the imagination only that we can form any conception of what are his sensations." Smith thought this empathy was widespread: "[T]his sentiment, like all the other original passions of human nature, is by no means confined to the virtuous or the humane, though they perhaps may feel it with the most exquisite sensibility. The greatest ruffian, the most hardened violator of the laws of society, is not altogether without it."¹⁰

Of course, empathy also has limits. Most of our thoughts are consumed with more immediate aspects of own well-being, with "hunger, thirst, the passion which unites the two sexes, and the dread of pain."¹¹ But the joys and sorrows that empathy brings are still *part* of our well-being, not some alien category. (Those who portray Smith as some arid economist who cannot imagine a vision of self-interest beyond "mo' money, mo' money" simply have not read him.) From our reasoning about how to attain that particular goal—how to alleviate the pain or increase the happiness of those imagined Others, how to understand the limits of our responsibilities to them—come our moral systems, our moral thinking. The spark of sympathy that leaps between our own eyes and the eyes of the person in pain, the smile that involuntarily comes to our lips as we imagine the reaction of a stranger to a thoughtful gift, this, according to Smith, is the root of Other-regarding morality.

Smith is not the only thinker to make an argument like this, of course, but was he right? Personally, I think he was—at least descriptively. I think our ability to imagine the situation of the Other—to "walk a mile in someone else's shoes," as Atticus says in *To Kill a Mockingbird*—starts the chain of moral reasoning, both for an individual and for a culture. To be sure,

we can then go off in very different directions. How best to vindicate this proto-moral concern? Through Kantian logic? Rule utilitarianism or act utilitarianism? Social welfare theory? We can build elaborate theoretical castles on this impulse, but I believe that original desire, that felt imperative, comes from the initial experience of sympathy, of empathy. It takes a sociopath, or narcissistic demagogue, to be without it.

Ascribing an important moral role to empathy might seem uncontroversial but it provokes justified skepticism among some moral philosophers. Empathy, they charge, is too blunt, innumerate, manipulable, and unreliable to be our guide to moral decision-making. We are more easily moved to empathy by those similar to us, leaving our moral vision clouded when it is most needed. Empathy provides no metric for moral decision-making in situations where there are scarce resources and many wrongs to right. (That is, always.) Instead, empathy's critics argue, we should focus on more rational measures of well-being, such as cost-benefit analysis or social welfare theory.

Finally, empathy cannot, by itself, resolve moral conflicts, nor should we think that ascribed personhood always dictates results. For example, one side of the abortion debate believes that empathy should make us stretch our definition of person to cover the nonviable fetus and perhaps even the just-fertilized embryo. Potential should be the warrant for personhood. The other side strongly disagrees and argues, in addition, that empathy should make us take more seriously the moral claims of women who do not believe the state has the right to "nationalize their wombs" in order to force them to carry a fetus to term against their wishes. My kidney might be the only hope of survival for someone with kidney disease. The suffering patient is clearly a person. We nevertheless resist the claim that the state has the right to compel me to provide my organs to sustain them. Personhood, in other words, is not the only issue, nor does empathy uniquely compel where we draw its lines.

These are powerful critiques, and I agree with some of them. But they miss the point of what I am doing here.

First, my goals are descriptive and predictive as well as normative and prescriptive. I ask how we will greet the emergence of synthetically created persons as well as how we should. A large part of both stories, I argue, will be whether it seems plausible to extend our empathy. The

critics of empathy agree that it plays a huge role in our moral deliberations; indeed, that is part of their criticism. Descriptively, then, they would have no complaints.

Second, even empathy's critics are not against all forms of empathy. Here is Paul Bloom, whose *Against Empathy* is perhaps the most comprehensive and full-blown critique:

But there is another sense of empathy or, to put it differently, another facet of empathy. There is the capacity to understand what's going on in other people's heads, to know what makes them tick, what gives them joy and pain, what they see as humiliating or ennobling. We're not talking here about me feeling your pain but rather about me understanding that you are in pain without necessarily experiencing any of it myself. Am I against this sort of "cognitive empathy" as well? I couldn't be. If you see morality in terms of the consequences of our actions—and everyone sees it this way, at least in part—then it follows that being a good moral agent requires an understanding of how people work. How can you ever make people happy if you have no idea what makes them happy? How can you avoid harming people if you don't know what causes them grief?¹²

Much of the empathy I describe is of exactly this kind. Finally, there is a missing step in the analysis. Bloom and others point out the irrational asymmetry of our moral reasoning: we focus more on the familiar and sympathetic, ignoring true need at a distance. True enough. But this presupposes that we see the issue as a moral one in the first place. We do not worry about my robot vacuum cleaner's moral claims or conduct a social welfare analysis of my toaster. They are machines. Before we can crank up our elaborate social welfare analysis or get our Kantian reasoning going, we need to be capable of imagining that there is even a moral issue to be considered. That is where synthetically created beings are likely to cause us problems.

As Smith points out, much depends on the initial act of imagination. Since our senses cannot give us the pains of others, our imagination must. But what if we do not think that "person" is in any way like us? What if we do not think they have any moral status at all? What if we would never even begin to conceive of putting ourselves in the shoes of someone of a different class, or a different gender, or a different nationality, or a different race or religion? Or of a nonhuman animal? After all, we have a history of doing exactly that. In such a case, their pains are no more real to us than is the pain experienced by a rock. Our imagination does

not make the leap, our empathy is never triggered, our morality is never at stake. How much more likely is that with synthetically created beings?

Can we be made to think otherwise? Can reason alone force us to expand or contract the realm of others whose travails we imaginatively inhabit? Sometimes. Logic is a powerful tool, at least for those who have learned that it offers insights as well as irritating challenges to deeply held beliefs. Our moral lives, and this book, are full of sloppy syllogisms: “You think X about Y, but not-X about Z. Yet Z is, in all relevant respects, actually a Y! Therefore, you are bound to think X about Z as well!” But if reason helps us step beyond the boundaries of our own flesh, sex, race, or religion, so too do narrative, imagination, and art.

If you look at the history of some social movement—for example, the long struggle to abolish slavery—you will find much moral, legal, and religious argument, but you will find those arguments resting on a base of “sympathy” that has been built up by telling stories again and again, stories that force us to put ourselves in the position of the Other. Moral philosophers sometimes downplay this portion of the history, as if it were simply a successful advertising campaign for a drug that scientific evidence later showed was good for you: the science does the real work, the ad just catches the eyeballs of fickle consumers. I think they misunderstand the process. Like climbers who brace themselves on alternating sides of a chimney as they ascend, we lever ourselves upward through both empathy-building narrative and dispassionate moral reasoning. Our moral tradition was built by both Spinoza and Shakespeare, Kant and Philip K. Dick. That is likely to prove as true with Hal and Chimp as it did with the moral debates of the past.

Betsy Clark, a friend and a brilliant historian who died tragically young, wrote a superb article chronicling this process in the abolitionist movement. *“The Sacred Rights of the Weak”: Pain, Sympathy, and the Culture of Individual Rights in Antebellum America*¹³ describes the explosive growth of antislavery sentiment in the northern United States in the period between the 1830s and the 1850s:

In 1835 an antislavery sympathizer leaving a lecture by Theodore Dwight Weld went home to dream that she was transported above the world; looking down at the United States, she saw “multitudes of sable figures, bending beneath a scorching sun—their backs lacerated by the whip—scourged, maimed, loaded

with irons—subject to every insult—and exposed to every gust of unbridled passions.” The dreamer, a Mrs. Sturges, drew from many discourses in describing her lengthy dream, but the fundamental trope of her visionary narrative was the story of the suffering slave, a trope that in the 1830s began to play a crucial role in an unfolding language of individual rights. Slaves had suffered for many generations by the time Mrs. Sturges had her vision, but in the 1830s their stories became newly audible and visible in the North, where graphic portrayals of slaves’ subjective experience of physical pain emerged as common antislavery fare. Augmented in the 1840s and 1850s by slave narratives and sentimental fiction, this genre, with its critique of interpersonal violence and sexual abuse, served as a vehicle for new arguments for a “right” to bodily integrity.¹⁴

The basic arguments against slavery had been around at least since the ancient Greeks. The issue was certainly alive in the early nineteenth century. Britain had actually criminalized the slave trade (though not slavery) in 1807, responding to the criticisms of abolitionists such as Samuel Romilly and William Wilberforce. What Clark describes, though, is a systematic, almost obsessive cataloguing of the horrific violence wreaked on slaves’ bodies, blow by blow and injury by injury. Narratives, both exhaustively factual and dramatically fictional, laid out the tiniest details of floggings, burnings, rapes, and brutalities—a process that culminated in Harriet Beecher Stowe’s 1852 novel *Uncle Tom’s Cabin*. The largely white, largely Christian audience responded with horror, indignation, and moral fervor. Sermons spoke of “the duty to feel an interest in the sufferings of others who are at a distance from us . . . to extend our sympathies beyond ‘the little limits of our state and our neighborhood.’”¹⁵ The fuel for that process of extensive sympathy was the laborious chronicle of the brutalities inflicted on the bodies of slaves, a chronicle that invited the white reader to switch places, to imagine those pains inflicted on his own tender flesh. It is on the ground of that constructed empathy that the moral argument against slavery then assumes its full force.

When we are not eyewitnesses to pain, we can only get access to it through someone else’s description. An account of the pain of others excites our sympathy “in proportion to the vivacity or dullness of the conception,”¹⁶ as Smith puts it, and this is true whether it is fiction or nonfiction. *Uncle Tom’s Cabin* would certainly qualify there. He goes on to describe how fiction can cause a suspension of disbelief, not just about the fact that it is merely a story but the fact that it is not about *us*.

Our joy for the deliverance of those heroes of tragedy or romance who interest us, is as sincere as our grief for their distress, and our fellow-feeling with their misery is not more real than that with their happiness. We enter into their gratitude towards those faithful friends who did not desert them in their difficulties; and we heartily go along with their resentment against those perfidious traitors who injured, abandoned, or deceived them. In every passion of which the mind of man is susceptible, the emotions of the by-stander always correspond to what, by bringing the case home to himself, he imagines should be the sentiments of the sufferer.¹⁷

It is one thing to cajole someone into extending their sympathy to other human beings. They are, after all, human. They feel pain. They are *just like us*. It is another to use fiction to do so beyond the species line, beyond the line of naturally occurring creatures altogether, to the android or the genetically engineered synthetic organism. In one sense, of course, it is the attempt to provoke the same imaginative, empathic leap that Smith and Clark describe. But how to induce that leap in the face of the visceral understanding that these beings are not like us, that they are synthetic and not natural? It is not merely that they were made. They were made by us. The earnest abolitionists could conjoin the sympathy for slaves' abused bodies with the Christian moral conviction that we are all God's children, endowed by our Creator with certain inalienable rights. But when we come to the android or the genetically engineered hybrid, we return to the argument I put into the mouth of the fictional creator of the Chimpys: "I am their creator, and I can assure you that I gave them no such rights." If you are a novelist or a filmmaker, how do you get past that objection? And does that effort tell us anything about the likely future of the debate over the personhood of artificial beings?

To answer both questions, I turn to *Do Androids Dream of Electric Sheep?*, a science fiction novel, and *Blade Runner*, the very different but equally brilliant movie based on it. Some of you will say that you hate science fiction. I would urge you to think twice. That is like saying you do not like books set in the past or books set in other countries. The generalization undermines itself as it is uttered. What you may hate is bad science fiction, and there is a lot of it. Strange, ugly words that play no role in the plot, lengthy descriptive passages about poorly rendered futures before any character does anything, societies that are technologically changed out of all recognition while sex roles are apparently stuck in the 1950s,

deus ex machina plot twists, with the emphasis on the *machina*; there is much to dislike here. But good science fiction, which, like a science experiment, changes just one or two variables about our world and then spins out a beautifully written story of the reality that ensues, characters like us in a world not ours, that is a thing of joy.

When Ursula K. Le Guin imagines a world without private property in *The Dispossessed*¹⁸ or Cory Doctorow conjures a society in which reputational capital is the real currency¹⁹—both worlds filled with sympathetic, flawed characters—they give us something precious: an ability to step away from our own world and find it, for a moment, strange. The Germans have a word (of course) for the sundering that happens when an author deliberately smashes the suspension of disbelief: *Verfremdung*. “It’s just a play,” screams the actor in a Brecht production, hoping to shock the audience out of the thrall of the theater and make them wonder if they need to do the same thing with the suspension of disbelief produced by the structures and roles of their own society.

For me, science fiction has always done this to my own quotidian world. It has done so even better than political or economic theory’s thought experiments (the Veil of Ignorance, the State of Nature, the Coase Theorem, the Efficient Capital Market), or the string of awful consequences a lawyer conjures up in an argument or a court decision (we call them “parades of horrors,” which sounds like a Diane Arbus Thanksgiving March). The ability to create a world and then be *limited* by it—to follow its dictates out to the end with rigor and discipline, while making strange the familiar—is no less to be prized in fiction than political theory. And that is what *Do Androids Dream* and *Blade Runner* bring to our discussion of the line.

A VOIGHT-KAMPFF TEST FOR HUMANS?

[T]he real question is not whether machines think *but whether men do*.

—B. F. Skinner, *Contingencies of Reinforcement*

Rick Deckard, the main character in *Do Androids Dream of Electric Sheep*, seems depressingly normal at first. He lives in a world recognizable in 1960s America. He is not quite the organization man, but he could play

him on TV. Cheery, cheesy commercialism pervades his life. Entertainment is provided through shows everyone watches and then discusses the next day at work. Corporations and their messages are everywhere. Social position is measured partly by the possession of certain status objects. Deckard's wife is depressed. He worries about his job. If he were a character of John Updike's or Philip Roth's, you'd expect him to have a mild midlife crisis punctuated by suburban adultery and martinis. It is as if the Civil Rights movement and the 1960s critique of the bourgeoisie never happened. Then again, in the real world from which I write these words, there are self-described white supremacists happily talking about the influence they recently had, and hope to have again, at the highest levels of the federal government. So how strange can an alternative reality be?

Strange, it turns out. There are a few minor adjustments necessary to get from our world to Deckard's. The novel is set in the United States after a nuclear war. The environment has been devastated. Millions are dead. In particular, nonhuman animals have been nearly eliminated. They are now treated with reverence; ownership of an animal is a potent status symbol. Deckard and his wife cannot afford one, so they keep up appearances and fool their neighbors with a robotic replica of a sheep while aspiring to upgrade to something real. Deckard works as a blade runner, a bounty hunter, trained to track down and kill androids—synthetically created robotic beings that do much of mankind's dangerous work, particularly off-planet, where most of humanity's best and brightest have already fled. These androids are so humanlike that a behavioral psychology exam—the Voight-Kampff Test—is needed to detect them. In a particularly dark moment of irony, it turns out that the test measures empathy, which, we are told, androids lack. Some of the questions, in fact, require showing an intensity of empathy for animals that the readers of this book might also lack. But the people in the almost animal-free world of *Do Androids Dream* are more reverent:

Rick, selecting question three, said, "You are given a calfskin wallet on your birthday." Both gauges immediately registered past the green and onto the red; the needles swung violently and then subsided. "I wouldn't accept it," Rachael said. "Also, I'd report the person who gave it to me to the police." After making a jot of notation Rick continued, turning to the eighth question of the

Voight-Kampff profile scale. “You have a little boy and he shows you his butterfly collection, including his killing jar.” “I’d take him to the doctor.” Rachael’s voice was low but firm. Again, the twin gauges registered, but this time not so far. He made a note of that, too.²⁰

Deckard worries about false positives with his Voight-Kampff Test. Perhaps a person with schizophrenia might show a replicant’s lack of empathy and accidentally be “retired”—note the euphemism for “eliminated”—by a blade runner. Look back at the test above, reader. Would *you* pass? Yet Deckard hardly ever worries about the converse. What if androids are in fact persons? Yes, in Deckard’s world they are artificially created, but what if they should still be recognized as people? What if it is a major failing of *human* empathy that they are not? Deckard’s society tests and then kills them based on a purportedly scientific measure of lack of empathy. And what precisely is that lack of empathy? That they have failed adequately to respond to a hypothetical test of ethics involving a nonhuman. Nothing could be more painfully ironic.

But that by no means exhausts the strangeness of this world. Take the Penfield mood organs that Deckard and his wife use, which allow one precisely to dial a particular emotional mood. This is something that goes beyond an attitude-adjusting beer after a hard day. This is cyberpunk Roth and Updike: “Run, neural code of Rabbit, run!”

Appearing beside him, her long nightgown trailing wispily, Iran shut off the TV set. “Okay, I give up; I’ll dial. Anything you want me to be; ecstatic sexual bliss—I feel so bad I’ll even endure that. What the hell. What difference does it make?” “I’ll dial for both of us,” Rick said, and led her back into the bedroom. There, at her console, he dialed 594: pleased acknowledgment of husband’s superior wisdom in all matters.²¹

To paraphrase the comedian John Oliver, #Irony. #Feminism.

Animals are loved and protected by law in Deckard’s world, far more so than in our own world. They are so revered that, because of their scarcity, many of them are actually replicas, which are cherished nonetheless. Almost perfect replicas of humans, however, are stalked and killed after being tested for their empathy for nonhuman animals. Moods, too, can be artificial, so that one is left doubting what the idea of authenticity even means. The replicants are violent. They murder several humans and attack Deckard. Does that show how important it is to protect the

boundaries of our species from homicidal murdering robots, or should we see it as the frenzied violence of a slave revolt, something that would not have happened if replicants had been treated as people? Deckard sleeps with one of the replicants. Does this prove that love can cross the line or merely that we have invented potentially homicidal sex dolls? There is a Kafkaesque scene—and I mean that in the “if you had lifted this and put it in *The Trial*, Kafka’s biographer Max Brod might have said, ‘Yeah, that checks out, leave it in’” sense—in which Deckard encounters an entire fake police station staffed by replicants. Maybe Deckard himself is a replicant? His partner? And if we don’t know who is a replicant and who is a human, how can we say they are less human than we are? These contradictions are wound through the plot. If you don’t look at them they are not obvious, but they are still unsettlingly visible out of the corner of the reader’s eye, a moral version of the graphic disorientation in an Escher drawing. Is this floor or ceiling? Up or down?

Philip K. Dick is playing with the line.

At every stage, the novel probes the coherence of our moral intuitions in a way that Adam Smith might have appreciated. Should empathy be the moral warrant for personhood? If so, does that prove that the replicants lack it, or that we do? If a synthetic entity can pass as human so that we cannot distinguish it from the real thing, does that suggest or require that we grant it personhood, and, if so, why? (Hal, the imaginary AI in the introduction to this book, was able not only to pass the Turing Test but even to administer it accurately to others, just as the blade runners do with their test. What follows from that fact?) Deckard’s society shows more empathy for nonhuman animals than our own. It is even able to suspend disbelief and cherish a synthetic replica of an animal. Does that prove that his culture is morally superior to ours, or is it simply a warning of the ease with which we can project qualities that do not exist onto a mere facsimile, as the colonel did with the mine-clearing robot? In a world of mood organs and electric sheep, what does the line between natural and synthetic even mean?

At the end of the book, the author seems to suggest that all foundational beliefs, whether in Mercerism—the empathetic religion of his world, which the book suggests is a fraud—or in the authenticity of his robotic sheep, are based on a willing embrace of delusion, a delusion that

might actually be worth that hug. The reader is left to wonder whether Deckard would apply the same logic to his own humanity. Or indeed, to ours. For where on earth could *our* consciousness come from, if there is no ghost in the machine? Alan Turing, the great computer scientist, made the same point about his Turing Test for machine consciousness. He makes use of an argument that B. F. Skinner and the behaviorists later developed into a successful intellectual franchise; since we do not have direct evidence of the mental states of other human beings, we could always solipsistically posit them to be rule-following automata:

I think that most of those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position. They will then probably be willing to accept our test. I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. 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.²²

Turing is trying to answer the question “can machines think?” What test will we set them in order to find out? If we set a higher bar than “seeming human,” can we meet it ourselves? Or, in the words of Skinner with which I began this section, “[T]he real question is not whether machines think but whether men do. The mystery which surrounds a thinking machine already surrounds a thinking man.”²³ Is the question not whether Deckard is a replicant but whether we *all* are? That question is one that *Blade Runner*, the movie based on *Do Androids Dream*, takes up.

Blade Runner has lots of similarities to *Do Androids Dream*, of course, but the differences might be more striking. The replicants are not cybernetic robots; rather, they are creatures of synthetic biology and genetic science, a very conscious choice by the director Ridley Scott, who was fascinated by the social changes that genetic engineering might bring. Are we dealing with Hal now, or Chimp? The film is set in dystopian Los Angeles rather than dystopian San Francisco. (Think this is a trivial difference? Ask a resident of either city.) Deckard (played by Harrison Ford) does not confront a mildly radioactive version of a 1950s organization man world but a landscape out of cyberpunk film noir, where darkness intertwines with occasional beams of light, natural or human-made, to dazzle, obscure, or highlight. When we first meet him, he is hunched

ineffectively against the rain (rain in LA!) and waiting to eat at an Asian noodle bar, White Dragon, where English seems to be the least-used tongue. He doesn't speak the bar owner's patois, Cityspeak (a foreigner in your own land!). He wants to order four dumplings, but each customer is allowed only two (consumers with money not able to eat as much as they want! In America!).

As he waits for his food, Deckard idly rubs his disposable wooden chopsticks against each other to get rid of the inevitable splinters after breaking them apart. An exchange student from Korea in my Law and Literature class said, "That's the most casually and unselfconsciously Asian thing I've ever seen a Western person do." Norms have changed, and if you have an implicit assumption that the majority of the United States is white and English-speaking, or that it doesn't rain in LA, the change might be disconcerting. Giant blimps float through the skies, displaying video advertisements for the off-world colonies that, with wonderfully jarring effect, juxtapose a cheery voice narration straight from a 1950s public health video, with a geisha-like female face in stylized makeup. Japanese brand names are everywhere. The movie was made at the height of the Asian-takeover fears of the 1980s and it shows. Magnificent corporate buildings rise, like Aztec pyramids, above the squalor of the streets. *Some* artificial persons are doing very well in this world, it seems. There are flying cars. And later, feral homeless children stealing machine parts *from* the flying cars.

Before Deckard's food even arrives, the viewer's sense of estrangement, of uncertainty, of identity crisis is well under way. All of that takes the director about a minute. It is a tour de force and one that, when watched again recently, was strangely prophetic about racial anxiety, xenophobia, and fear of the Other. When we find out that Deckard's job has been to police the boundary line of our species, it all just fits right in. "We want him on that wall!" Or do we?

Like Dick, Ridley Scott produces disorientation and sudden flashes of enlightenment in a flickering, moral seizure-inducing pattern. The very beginning of the film features a replicant named Leon (played by Brion James) being given the Voight-Kampff Test by a blade runner. Leon is like the student who fights the hypothetical question in a classroom discussion of ethics: the student who responds to the trolley problem by saying

that he always takes the subway, and why are those kids on the track in the first place? Leon constantly pushes back at the questions he is asked. Told by the interviewer that he is in a desert, he wants to know which one. Given a hypothetical situation involving him flipping over a tortoise and leaving it in the hot sun, he keeps asking questions. “What’s a tortoise? Why would I flip it over? Did you think up these questions or did someone else write them for you?”

We are caught between wondering whether Leon is “on the spectrum” and realizing that we have so internalized the role of student in the highly artificial world of test-taking, with its constraint-free questions designed to probe certain responses, that Leon’s perfectly normal inquiries seem naive. His leaden puzzlement is actually endearing—right until the moment when the interviewer asks about Leon’s mother. “Let me tell you about my mother,” says Leon. Then he pulls out a gun and shoots the interviewer. Motherhood, it seems, is a touchy issue for the synthetically created. Of course, in shooting a member of the species that has created him, who is actually employed to track him down and kill him, maybe Leon *is* telling us about his mother. There is an instantaneous shift from earnest, confused student asking for reassurance from the teacher to homicidal killing machine striking back at its creators. Scott is not going to make it easy for us by making the replicants warm and fuzzy.

The Nexus 6 replicants at the heart of the movie have escaped back down to earth by stealing a shuttle and killing the crew. Led by Roy Batty (played by Rutger Hauer), they are in search of ways to prolong their very short lifespan—a limit hardwired into their DNA by their creator, the Tyrell Corporation. Psalm 90 tells us that the Lord has given us “three score years and ten.” Tyrell’s creations get four years.

Their search for a way to stave off impending death gives the movie poignancy amid the menace, with strangely touching moments. Roy and Leon question one genetic designer, Hannibal Chew, who disclaims any knowledge of biological lifespan. “I just do eyes.” The replicants are standing in street clothes, quite comfortable in a cryogenically chilled facility as the designer shivers in front of them. Despite his entirely warranted fear of what is about to happen to him, Chew says, “You Nexus, huh? I design your eyes.” It is not entirely clear, but it looks as though he makes an abortive gesture toward the face of a being who is doubtless

about to kill him. It's an odd moment of craftsman's pride, even tenderness. "Chew. If only you could see what *I* have seen, with *your* eyes," replies Roy. Yes, Ridley Scott is telling us, this is a designed creature and a dangerous one, confronting its makers in anger. But it is also an "I" with emotions and memories, a person inside that skull, who wishes that the designer could see the world through *his* eyes, walk a mile in *his* shoes. Of course, we cannot change places with the Other, except in our imagination. Will we? It is Adam Smith's discussion of empathy all over again. And the question I posed earlier—how can a being we *designed* also be a person?—is front and center.

PRIMING: THE MORAL STROBOSCOPE

Philip K. Dick's *Do Androids Dream of Electric Sheep* uses a number of mental dislocations to shake our brains out of their familiar patterns, to confront the Other with an innocent eye. There is the Voight-Kampff Test that denies interviewees personhood if they cannot feel enough empathy for a different species and yet never causes the humans to doubt their own lack of empathy for *their* creations. There is almost fetishistic worship of nonhuman animals in an ecologically ravaged world. The book brilliantly uses the power of language over our imagination to run thought experiments. What if there were a mood engine that precisely and artificially changed moods? What if there were a religion based on empathy? But *Blade Runner* is a movie. It can show us the line rather than just tell us about it.

This book is about the line of personhood. What is on the other side of that line? What are the edge cases, the things that we use to demonstrate the boundary between us—persons, legally recognized entities with an array of rights—and nonpersons? It is easy to say that a chair or a table is not a person, but what about the closer calls, the examples that philosophers through history have used to support their definitions of the "it" that makes us, us? Most obviously, we have *nonhuman animals*. They are like us in many ways, but whereas I can own a chimpanzee or a dolphin, the reverse is not the case. We have simulacra: the mannequin, statue, or wax model. They look eerily human, but we know they are not. We can even add functional similarity to physical similarity. We have robots that

look and act like humans, and we have software programs that mimic aspects of human speech or cognition. We have artificial, genetically engineered beings based in part on human genetic material. Finally, we have the temporal dimension of the line: dust to dust, ashes to ashes. When does *something* become *someone* and vice versa? Does life, does personhood, begin at conception, viability, birth? Does it cease when there is no breathing, no heartbeat, no brainstem activity? Time, too, is an edge case. Animals, mannequins, robots, software emulations of humanity, and the life-death divide. Each of these might help us illuminate what makes us, us. *Blade Runner* explores all of them.

J. F. Sebastian is one of my favorite characters in *Blade Runner*. A talented genetic designer who suffers from Methuselah Syndrome, he is aging too fast and will die young, yet his is a natural condition and not the programmed lifespan of the replicants. He lives alone in the decayed and abandoned Bradbury Building, an 1893 Los Angeles landmark whose name also ironically calls to mind a science fiction legend, though nothing could be less like the endless childhood summers of Ray Bradbury's fiction than the dark, rainy, and dilapidated world of *Blade Runner*. His only companions are an array of fantastical mannequins and toy robots or genetically engineered play-animals—scaled-up versions of a child's clockwork soldiers or stuffed teddy bears. The stiffly marching, Pinocchio-nosed Kaiser and his companion bear suggest a Victorian playroom but also a twenty-first-century robotics shop or genetic laboratory. The marching figures' movements are jerky, like clockwork, but they speak, see, and even greet their master by name when he returns home. It is both beautiful and sad. As Sebastian himself points out in explaining why he is not lonely, "I *make* friends. I am a genetic designer." Perhaps his glorious toy room is a reflection of, or a comfort for, his lost youth. He is a sympathetic character. The toys do not seem fully sentient but he is literally *making* friends, and, in his day job, he has been one of the designers of the replicants.

One of those replicants, Pris (played by Daryl Hannah), is sent to befriend Sebastian as part of the attempt to lengthen their four-year lifespan. Pris is a "pleasure model" replicant. If humans can design sentient, genetically engineered beings, the movie suggests, nothing is less surprising than turning them into sex toys. Indeed, today's tech journalism has

featured many stories about the use of robots as companions for those who live alone²⁴ and, yes, as talking sex dolls.²⁵ Siri in lingerie, as it were. As an article in the *New York Times* puts it:

[C]onsenting is not something these robots are capable of. That doesn't matter, the argument goes, because these are not women, but animatronic objects, so consent is not necessary. The same reasoning is used to deflect fears that such robots could influence societal attitudes toward women. "She's not a someone. She is a machine," their creators are quick to respond when questions of moral ambiguity are raised. "Is it ethically dubious to force my toaster to make my toast?"²⁶

Pris is dressed in an outfit straight from a 1980s punk club: racoon-like eye makeup, torn fishnet stockings, boots, a wild, platinum blonde thatch of hair. She conceals herself in a pile of trash near Sebastian's building, like a Dickensian street child, then bursts out in fake alarm when he arrives, colliding with him in the process. She is alone, "kind of an orphan"—which, as a replicant, she actually is. She is beautiful, seemingly vulnerable, and has nowhere to go. Sebastian offers her shelter.

Psychologists have explored the power of priming a viewer or experimental subject, providing a context that will cause them to interpret or remember material in a particular way. Some of that research was speculative and made implausibly grand claims that have been impossible to replicate. However, the basic mechanism has been repeatedly tested and confirmed. To quote *Psychology Today*:

Priming is a nonconscious form of human memory concerned with perceptual identification of words and objects. It refers to activating particular representations or associations in memory just before carrying out an action or task. For example, a person who sees the word "yellow" will be slightly faster to recognize the word "banana." This happens because yellow and banana are closely associated in memory. Additionally, priming can also refer to a technique in psychology used to train a person's memory in both positive and negative ways.²⁷

Throughout the scene in Sebastian's apartment, the movie sends a stroboscopic set of images designed to prime us to see Pris, and later Roy, as different entities on either side of the line. It is done with malice aforethought and at remarkably high speed. Pris's raccoon eye makeup does remind me fondly of punk rock clubs in the 1980s, but it also looks like, well, an actual raccoon, particularly when Pris bends and sniffs repeatedly

at the sleeping Sebastian. Humans do not normally sniff at each other like dogs. She is an animal! Pris is almost inhumanly perfect both in looks and physical abilities. When she reaches into a glass beaker of vigorously boiling water to pull out a hard-boiled egg, without even noticing the heat, we are unsurprised. She tosses the egg to Sebastian who has to juggle it because it is so hot. She is a robot! She is beautiful and beguiling: more than one audience member has looked at her longingly and then had the cognitive dissonance Ridley Scott surely intended. Are you admiring a beautiful woman or looking lustfully at a sex doll? Ew. When Roy comes to visit, they hungrily kiss each other, showing no shyness in front of Sebastian, but they also sniff at each other like two dogs reunited. She's an animal! Or a sex toy! Or in love! But Pris is also a child. She shows joy at Sebastian's playful creations and fear of pursuit, and she oscillates between manipulation and apparent affection for Sebastian himself. Hearing the news of Leon's death from Roy and realizing that there are only two of them left, she is distraught: "Then we are stupid, and we'll die."

When Deckard arrives looking for her, she hides among Sebastian's clockwork figures, draped in gauze. The camouflage is remarkably successful. Her immobility, her waxy perfection of features and form—she is a mannequin! A doll! Right until Deckard lifts the gauze veil with the barrel of a gun, and, shrieking like a hawk, she attacks him with inhuman strength. She is a killer android! Daryl Hannah is a former ballet dancer and performed some of her own gymnastic stunts. In the middle of the fight with Deckard, she performs an absurdly difficult gymnastic tumbling routine—intentionally jarring coming from a homicidal killing machine. *The Terminator* meets Olympic floor exercises. When Deckard finally shoots her, brutally interrupting the perfection of her flips and handsprings, her body has a violent seizure on the ground as if she were being electrocuted, while she shrieks like a dying animal. And bleeds. The priming comes fast. Flash. A beautiful woman. Flash. A killer android. Flash. A lifelike mannequin. Flash. A child. Flash. A dying animal, screeching in unbearable pain. For me, that moral stroboscope explains the power of the movie better than any other factor. It is a remarkable piece of work and a deeply troubling one. Is it really so easy to

manipulate our sympathies? Are our categories so unstable? Do we have too much empathy? Too little?

Blade Runner is not alone in exploring these themes, of course. Contemporary depictions of artificial humanity such as the HBO television adaptation of Michael Crichton's 1973 *Westworld*²⁸ touch on many of the same issues. The robotic hosts in the Western theme park in that show become a canvas on which humans can paint our own moral portrait, revealing what we would do to those who had no rights but looked like us. Rape, murder, torture, a few rescue fantasies: the picture is a depressing one. If the moneylender in Dostoyevsky's *Crime and Punishment* had been described to Raskolnikov as a mere clockwork toy, think how his murderous Nietzschean fantasies could have flourished. But perhaps *Blade Runner's* replicants disorient us about the line in a different, or additional, way. *Westworld* gives us a dawning realization—"Wow, the hosts are sentient! And they are in revolt. (Dolores seemed so nice before!)" *Blade Runner* wants to disorient us from moment to moment and scene to scene, snapping us back and forth through a range of implicit characterizations: animal, android, psychopathic monster, person, sex doll, mannequin. Return for a moment to the Adam Smith excerpt on sympathy with which I began this discussion: "Though our brother is on the rack, as long as we ourselves are at our ease, our senses will never inform us of what he suffers. They never did, and never can, carry us beyond our own person, and it is by the imagination only that we can form any conception of what are his sensations."²⁹

But what if our imagination flipped back and forth between viewing him as our brother, a mannequin, an animal, an implacable killer android, a toaster that can't say no? A person? *Blade Runner* seems to tell us that our empathy—already unreliable even in valuing others of our own species—will be sorely challenged in two ways. First, the possibility for priming will be more present than in any of the prior personhood wars because many of the primings will be *true*. Chimpy *is* partly an animal. Hal's consciousness *is*, in part, a result of human programming. Pris *is*, in part, a synthetic construct designed as a sex toy. Roy *is* a frightening android soldier. All of them *are* creatures of our own design, our own creation.

This brings us to the inscrutability paradox. If we know how you were designed, if we understand how you “think” because we programmed you to do so, how can we see you as anything but a mannequin or a wind-up doll rather than an entity with free will? How does empathy work in that situation, particularly when there will be many who have economic or ideological interests in pushing the particular priming that humanizes or dehumanizes our creations? Particularly when our skepticism about each new claim of personhood might have very good arguments behind it? But if your behavior is produced through methods and paths we do not completely understand, which is already the case even in some of today’s neural nets, then will we be caught between fear and incomprehension? Neither of those is a fertile basis for empathy.

Second, our empathy will be challenged because *we can design around it*. These are our creations. If a particular body shape or set of responses or pattern of speech or neotenous facial features make us class an entity as human, then the designers can choose to keep it or change it. What would Smith make of a world where our brother could be designed to look like he was *part* of the rack? Or a cartoon character getting a nice lumbar stretch? Smith dealt with a world where the differences across which empathy must reach were naturally occurring, even a given. In the strange world of Hal and Chimpy, those differences will be *chosen*.

In *Blade Runner’s* climactic moments, Roy uses a ruse to gain entrance to the Tyrell Corporation headquarters and to Tyrell’s own rooms. The creation confronts his creator. Tyrell asks why he has not come before. “It is not easy to meet one’s Maker,” says Roy. A moment later comes one of my favorite pieces of dialogue. Roy asks whether the maker can change his own creation. “What seems to be the problem?” says Tyrell, like an urbane physician. “Death,” answers Roy. Can Tyrell help? The answer, it turns out, is no, but Tyrell tells him to revel in his time. “The candle that burns twice as bright burns half as long, and you have burned so very, very bright, Roy.” As a consolation, it is dramatically unsuccessful. Though Roy seems to be about to confess to his father—“I have done questionable things”—he then adds, “but nothing the god of biomechanics wouldn’t let you in heaven for.” And on that note, Roy passionately kisses, and then kills, the man who made him. Tyrell, too, has done “questionable things” unrestrained by any god of

biomechanics. The creation is judging the creator. How will we be judged by ours?

Readers find many, many messages in *Do Androids Dream* and in *Blade Runner*—one reason for the enduring power of those works. Clearly, they are both meditations on empathy and personhood, but to me it seems they go beyond a simple injunction that we should love (electronic and genetically altered) others as ourselves. Instead, I see two themes, both central to this book.

First, our beliefs about naturalness, about identity, about empathy, are built on assumptions that start to seem arbitrary, even ludicrous, when presented in the fun-house mirror of an alternative world that is recognizable yet different from our own. We could mock the Voight-Kampff Test for its ironic focus on empathy for animals to deny empathy to androids. What hypocrites and fools the inhabitants of that world are! How blind they are to their own contradictions! But that is not the message I get from *Do Androids Dream* and *Blade Runner*. It is more like “judge not, lest we be judged.” Do you think we will do better? That question should prompt worried humility rather than hubristic condemnation.

The process of empathy that Adam Smith describes may be a basis for ethics, but it will be based on a leap of faith, a projection of identity that will probably rest on a pattern of beliefs full of blind spots and inconsistencies, one that will be in productive tension with our moral theories and our attempts to reason our way to the right answer. Each side will need the other. Our discussions of synthetic personhood will exist in a dialectic between sympathy and syllogism, leaps of empathy and flashes of disgust, hopes that we can realize the angels of our better nature and deep fears that our creations will destroy us. Or judge us. And, as I pointed out earlier, our empathy will be subject to radically conflicting, and partially truthful, primings that cause us to humanize or dehumanize our creations—to push them away into the realm of nonhuman animal, clockwork mannequin, clever software emulation, toaster that can't say no or, indeed, to recognize them as persons though we know that we made them. What's more, some of those primings will be designed

into our synthetic creations *on purpose*. Think of the Chimpy deliberately designed to look ape-like and to lack spoken language. Humanities graduate students quickly learn that if they do not know the answer to a question, “it is socially constructed” will generally get them off the hook. Our process of empathy for our synthetic creations will exist in a landscape that is literally made by our own hands. This is “social construction” with a vengeance.

Second, uncertainty about personhood does not stop when we get to our own species line, or our own identity. Deckard is not the only one to doubt his humanity. When we come to explain why *we* are conscious, human, how it is that there is a thinking “I” inside of humanity that has been given dominion over the animal and mechanical world, we could be subject to the same skepticism that the Google engineers directed at Hal. The question is not whether Deckard is a replicant. The question is whether we all are.

Earlier I quoted Turing pointing out that it was hard to prove that humans, too, were more than rule-following automata. He was not the first person to make this argument. In 1887, Samuel Butler had made the same point: “[T]he theory that living beings are conscious machines, can be fought as much and just as little as the theory that machines are unconscious living beings; everything that goes to prove either of these propositions goes just as well to prove the other also.”³⁰ It is (electric) turtles all the way down. To solve the problem of recognizing the Other, it seems, *Do Androids Dream* and *Blade Runner* are telling us we must first know ourselves.

In the movie’s final scene, Roy returns to Sebastian’s apartment alone; we are left to wonder what he has done with the gentle man who helped him and Pris and who tried to flee when he saw Tyrell killed. Roy finds Pris dead, calls her name, cries, daubs himself with her blood, and, howling like a wolf, sets off in pursuit of Deckard. The stroboscope of primings begins again, flashing like the flickering lights in which the scene is filmed. Roy goes from chanting murderous nursery rhymes, to animalistic howling, to inhuman feats of strength, to Socratic humor, toying with Deckard like a cat with a mouse. “Not very sporting to fire on an unarmed opponent. I thought you were supposed to be good. Aren’t you the . . . *good* man?” Is he?

As the chase continues, we can see that Roy's death is coming closer. His hand seizes up and—metaphor alert—he jams a nail through his palm to keep it open. Yet if Roy is supposed to be a Christ figure, the gift of the god of biomechanics is a horrifyingly ambiguous one, and it is not clear our sins will be expiated by his death.

Finally, Roy has a terrified Deckard defenseless, dangling by one hand from the roof's edge and about to fall. Deckard's fingers slip. Inexplicably, Roy *reaches out*, across that short but enormous divide, grasps Deckard's wrist, and saves him. He gives mercy to the man who has killed his lover when he himself has only moments of life remaining. In those moments, Roy returns to the same point he had raised with Hannibal Chew, the things his eyes have viewed. "I've seen things you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I watched C-beams glitter in the dark near the Tannhäuser Gate. All those moments will be lost in time, like tears in rain. Time to die."

These are profoundly *human* qualities: mercy; the bittersweet confrontation with mortality; the transitory nature of our consciousness and our memory; the impossibility of truly knowing another, of feeling what he or she feels or has lived. The essentially solitary way that all of us, not just the replicants, confront our impending crossing of the line between life and death, person and thing. "At the narrow passage, there is no brother, no friend."³¹ And yet, there is a consolation: our ability through language and art and empathy and dark, dark humor to share something, to reach out across that great divide.

Do we want to accept the gift of Roy's mercy—sparingly dispensed to be sure—but no less remarkable for it? Are we in fact willing to believe what his eyes have seen, to share his memories? Or do we leave him on the other side of the line? All of that, it seems, depends on how we define us as much as how we define him.

Perhaps that is the most important thing to realize from this chapter. Grappling with the question of synthetic Others may bring about a reexamination of the nature of human identity and consciousness that is unparalleled since secular philosophers declared that we would have to learn to live with a God-shaped hole at the center of our world. To draw the line for our creations, we must first draw it for ourselves. We have our own Voight-Kampff Test to face.