Reflections in a Computed Eye: Ethical Questions about Artificial Entities

Judith Donath
Berkman-Klein Center, Harvard University

Introduction

This chapter is about the ethics of our relationships with artificial entities—bots, robots, and other computational systems created to interact with us as if they were sentient and autonomous individuals. They may be embodied as robots or exist only in software; some are clearly artificial while others are indistinguishable, at least under certain conditions, from human beings. When are such interactions helpful or harmful? How do our relationships with computational entities change our relationships with other human beings? When does it matter if we interact with a machine or a human, and why?

Sentience—the ability to have emotions, to feel pain and want to avoid it—is a core concept here. We have ethical responsibilities to sentient beings that we do not have to nonsentient objects: it is cruel to kick a dog, but not a rock. While actually sentient artificial entities might someday exist, they are as yet only a theoretical possibility. All currently existing artificial entities are nonsentient, but—unlike a rock—their interactions and designs evoke the impression of conscious entities with personalities and emotions.

Simulated sentience is the primary focus of this chapter, highlighting our relationship with entities that appear to be sentient but are not. Some are quite simple; our tendency toward anthropomorphism can make the output of even primitive programs appear to us as the behavior of a cognizant mind. Others are impenetrably complex, with sophisticated imitations of conscious and intelligent

behavior that are nearly impossible to distinguish from the actions of an actually conscious being.

Some of the ethical issues we will examine involve our personal relationships with artificial entities. People seek companionship from artificial assistants, hold funeral services for broken robot dogs, and confide in simulated therapists. The relationships that some warn are a threat to humaneness, if not to humanity, are proving to be quite popular. Under what circumstances are they helpful or harmful? How do such human/machine interactions affect our relationships with other people? How does the machine performance of emotion differ from human impression management or from the inauthentic expression required by, for example, the service industry? When and why does it matter that the other does not actually think? The key issues here concern empathy and the function that caring what others think plays in society.

We will also address ethical issues in the design and deployment of artificial entities. In their mimicry of sentient beings, artificial entities are inherently deceptive: even one that types "I am a bot" implies, with its first-person pronoun, a self-conscious being. And many artificial entities are designed to be as persuasive as possible, eliciting affection and trust with features such as big childlike eyes and imitative gestures. Some are made with beneficial goals—to serve the user as teacher, wellness coach, etc.—but these same persuasive techniques can manipulate us for harmful and exploitive ends. What are the ethical responsibilities of researchers and designers?

While some artificial entities attempt to pass as human, many are clearly robots or software agents; the illusion they project is of a sentient but also distinctly artificial being. Yet the popular vision of truly sentient machine beings is generally foreboding—they are often portrayed as a potent, if not the final, enemy of humanity. Why do we see this future so darkly? While understanding the ethical issues surrounding our relationship with artificial entities is important in itself as social robots and software agents become increasingly present in our everyday lives, these queries also shed revealing light on our relationships with each other and with other living things.

Scope and Definitions

We will start with some definitions. Much discussion about today's nonsentient social robots and programs uses language that implies they have feelings and intentions, blurring the important distinction between "X is a robot that feels" and "X is a robot designed to appear as if it feels." Having a clear understanding of what is meant by intelligence, sentience, and consciousness and using them precisely is important for many ethical considerations.

Intelligence is often described as the ability to learn and apply knowledge or to solve complex problems.[1] It is an observable property defined by behavior—finding clever solutions, acting resourcefully. Thought of this way, we see a migrating bird, an insect-hunting bat, and a theorem-proving human as problem solvers each of whom require considerable, albeit very different forms of, intelligence. Thought of this way, we can easily refer to a machine as intelligent if it solves difficult problems. In this usage, the internal state that produces the intelligent behavior does not matter.

Yet intelligence is not a precisely defined term.[2] It is sometimes conceptualized as an inner quality, as when we say the migrating bird is not really intelligent, but is just acting on instinct. Computer scientists joke that use of the term "artificial intelligence" also reflects this enigmatic property: computer programs that solve complex problems using methods we do not understand are "artificial intelligence"; when we do understand them they are "algorithms."

Sentience is the ability to experience sensations and emotions: to feel pain and pleasure, and to want less of the former and more of the latter. A nonsentient creature may move away from certain things and toward others, and even have a suite of behaviors that aid its survival and reproduction, but it is not motivated to do anything: it simply exists. With sentience comes motivation: a creature that

---

[1] Tegmark, Max. "Let's Aspire to More Than Making Ourselves Obsolete." In *Possible Minds: Twenty-Five Ways of Looking at AI,* edited by John Brockman. (New York: Penguin, 2019) 76-87.

[2] Legg, Shane, and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence." *Minds and Machines* 17, no. 4 (2007): 391-444.

experiences certain sensory inputs as painful will want to avoid those; it will want to repeat pleasant ones. Sentience is now believed to be the foundation of learning, which gives sentient creatures much greater flexibility in their relationship with the world.[3]

Sentience is central to ethics because we have responsibilities toward sentient beings that we do not have toward, say, a rock.[4] Most people would agree that we should not inflict needless pain on something capable of experiencing distress. However, which beings are included in that category and what to do when that responsibility conflicts with other needs and desires are highly contested questions.

The term *conscious* refers to sentient beings that are self-aware—that have a sense of purpose and of themselves as individuals in the world. The term can be fuzzy: there is no clear behavioral marker of consciousness nor even an agreed-upon description of the internal experience. Historically, the rationalist, Enlightenment view was that consciousness was the affectless mental acquisition and manipulation of a symbolic representation of the world. Some believed that it required language and thus humans were the only conscious animal. Today, consciousness is increasingly understood to have evolved through social interaction, beginning with the bonding of parent and offspring; it is built on the emotional scaffolding of sentience.[5] And ethological and neuroscientific studies affirm that humans are far from being the only conscious animal: many mammals,

---

[3] Bronfman, Zohar Z., Simona Ginsburg, and Eva Jablonka. "The Transition to Minimal Consciousness through the Evolution of Associative Learning." *Frontiers in Psychology* 7 (2016): 1954

[4] Broom, Donald M. *Sentience and Animal Welfare* (Wallingford, UK: CABI, 2014); Singer, Peter. *Practical Ethics* (Cambridge, UK: Cambridge University Press, 2011).

5 Singer, Tania, Ben Seymour, John O'Doherty, Holger Kaube, Raymond J. Dolan, and Chris D. Frith. "Empathy for Pain Involves the Affective but Not Sensory Components of Pain." *Science* 303, no. 5661 (2004): 1157-1162.

birds, even cephalopods are aware of themselves and others and move through life with intentions.[6]

These differing views of what consciousness is have important repercussions for ethics and AI. In the classical view—which remains influential in some AI research as well as popular belief—consciousness is closely entwined with intelligence, the acquisition of knowledge, and problem solving. This contrasts sharply with the biological view, supported by current research, that consciousness is fundamentally social and emotional, having evolved from simple sentience as creatures began to bond and care for each other.

Consciousness is important in ethics because the basis of morality is here, in the evolution of traits such as attachment, empathy, and the desire for justice and social order. To care about how one is perceived by others and about one's effect on them—concerns available to the conscious mind—is arguably the very foundation of ethics.

Both sentience and consciousness are inherently private experiences. We cannot directly experience what it is like to be another being—human, animal, or robot. Our assessment of what it is like to be another, including what, if anything, they feel, is based on external and perceivable appearance and behavior. I assume other people are conscious because I know that I am conscious and we are biologically and behaviorally similar; it is, however, an assumption and not direct knowledge.

As we look at other species (or artificial entities), we make inferences about what it is like to be them—what their internal experience is—by analogy. The more something resembles ourselves, the more we assume his, her, or its experience to be similar to our own. This rule of thumb has led us to vastly underestimate the cognitive ability and sensate experience of many nonhuman

---

[6] Thompson, Evan. "Empathy and Consciousness." *Journal of Consciousness Studies* 8, no. 5-6 (2001): 1-32; Panksepp, Jaak. "Affective Consciousness: Core Emotional Feelings in Animals and Humans." *Consciousness and Cognition* 14, no. 1 (2005): 30-80; Peter Godfrey-Smith, *Other Minds: The Octopus and the Evolution of Intelligent Life* (London: William Collins, 2016).

animals and, as we shall see, to overestimate the capabilities of bots and other nonsentient human inventions.

Precursors: Turing and Weizenbaum

Our inability to directly observe the experience of being another is the problem at the core of Alan Turning's 1950 paper, "Computing Machinery and Intelligence," that marks the beginning of the field of artificial intelligence.[7] Turing introduced the paper by saying, "I propose to consider the question, 'Can machines think?'" and then immediately rejected the question on the basis that the words "machine" and "think" were too vague and limited by everyday experience.

Instead, he proposed a test, the Imitation Game, now popularly known as the Turing Test, which he argued was a "more accurate form of the question." In this test a human judge chats (via text) with two hidden contestants. Both claim to be human, though only one is—the other is a machine. The judge is tasked with determining which one is telling the truth. A machine that can consistently pass as human, Turing argued, should be considered intelligent.

It is a peculiar article and a hugely influential one.[8] It anointed deceptively passing as human as the key goal—or even as the definition of—artificial intelligence. And it deftly limited the domain in which this goal needed to be achieved to text-only communication.

Turing famously predicted that in fifty years computers would have reached the point that they would be consistently able to fool a human judge.[9] But he also made a second prediction: that by the time hat computers could pass as human, our use of language would have changed significantly. He said, "The original

---

[77] Turing, Alan. "Computing Machinery and Intelligence." *Mind* 49 (1950): 433-460.

[8] As a philosophical article, it is odd. It has pages of discussion about the nature of a digital computer but the central argument, that the Imitation Game is a satisfactory substitution for the question of whether machines can think, is rather glossed over.

[9] Specifically, that they would be able to "play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning."

question, 'Can machines think' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."[10] Though this second prediction, about the change in our culture and the meaning of words, is less noted, it was prescient. It is through such changes in language—in how we speak about thinking, about machines wanting and liking things—that our culture and ethics evolve.

About fifteen years after "Computing Machinery and Intelligence" was published, Joseph Weizenbaum created the first program capable of carrying on such a text conversation. He named this program ELIZA, after the character in George Bernard Shaw's play *Pygmalion* who "learns to speak increasingly well."[11] Weizenbaum's research goal was to interact with computers using natural language; with this project he sought to show that a simple sentence-parsing program with some semantic heuristics could carry on a coherent conversation . ELIZA was able to find the topic of a sentence and had rules for forming a response, but had no contextual information about the world.

It was an approach quite different from what Turing envisioned. Turing's belief in the significance of carrying on a humanlike conversation was not as shallow an assumption as it seems now. He described a potentially winning machine as having processing power equivalent to the human brain (though he quite underestimated the human brain's complexity and power); it would initially be programmed to simulate an infant and would then be taught, much as a child is. Turing's views about the brain, learning, and children are remarkably naive. But the key point is that he believed that a machine that would pass his test would be one that was imbued with a mind analogous to that of humans, able to learn, to reason. Furthermore, though Turing remained adamant that we rely solely on external behavior in judging what is thinking, he outlined the possibility of a state

---

[10] Turing, Alan. "Computing Machinery and Intelligence." *Mind* 49 (1950): 442.

[11] Weizenbaum, Joseph. "Contextual Understanding by Computers." *Communications of the ACM* 10, 8 (1967): 474.

change, analogous to the critical mass of an atomic reaction, that would mark a qualitative leap in mental ability and creativity.

ELIZA succeeded in sustaining conversation not through sophisticated technology but through, somewhat inadvertently, exploiting the way people make sense of each other. ELIZA was designed to respond based on scripts that would encode conversational rules for different roles. The first and by far most famous script Weizenbaum made for ELIZA was DOCTOR, modeled after a "Rogerian psychologist." His choice of this therapeutic framework was pragmatic: "the psychiatric interview is one of the few examples of categorized dyadic natural language communication in which one of the participating pair is free to assume the pose of knowing almost nothing of the real world."[12]

People were entranced with the computational "therapist." Even Weizenbaum's secretary, who knew the scope and point of the work, said upon trying it out that she wanted to chat with it further--in private.[13] Others took seriously the notion of the computational chat-bot as therapist, one that would be available to all, inexpensive and tireless.[14] At first Weizenbaum assumed this enthusiasm, which he judged to be misplaced, was due to the novelty of the interaction; future iterations should and would be designed to eliminate the "illusion of understanding."[15]

Weizenbaum's responses over the years show his growing alarm at this response. The quick willingness to accept a text-parsing program as an entity worthy of relating to, a repository for one's confidences, became to him an indicator of a deeply disturbing lack of concern about the humanity of the other—

---

[12] Weizenbaum, Joseph. "Contextual Understanding by Computers." *Communications of the ACM* 10, 8 (1967): 474-480.

[13] Weizenbaum, Joseph. "Contextual Understanding by Computers." *Communications of the ACM* 10, 8 (1967): 474-480.

[14] Kenneth M. Colby, James B. Watt, and John P. Gilbert, "A Computer Method of Psychotherapy: Preliminary Communication," *Journal of Nervous and Mental Disease* 142, 2 (1966): 148-152.

[15] Weizenbaum, Joseph. "Eliza—a Computer Program for the Study of Natural Language Communication between Man and Machine." *Communications of the ACM* 9, no. 1 (1966): 43.

a lack of empathy and of even any interest in the mind and soul of the other. Weizenbaum had come to America fleeing Hitler's Europe and knew vividly and with horror the devastating effects of dehumanizing other people. He spent much of the rest of his career warning about the danger computation posed to society. Turing argued that we need to accept intelligent behavior (which he had redefined as the ability to convincingly imitate a human in a text conversation) as sufficient evidence of machine thinking. Fifteen years later, Weizenbaum's ELIZA, a clearly non-thinking, sentence-parsing chat-bot, posed a counterexample by demonstrating how easily the illusion of intelligence can be made. Dismayed by people's enthusiastic embrace of ELIZA's therapeutic potential (and computers in general), Weizenbaum came to believe that the willingness to accept machines in such roles was a significant threat to humane society. These positions, taken in the earliest years of AI research, delineate the big ethical questions surrounding artificial entities and provide the starting point for our analysis.

Where Are We Now?

Turing's prediction—that in limited conversations, machines would be indistinguishable from humans—was off by a few years. In 2000, there were no computers that were able to consistently pass as human after five minutes of text-based interaction. But a couple of decades later his prophesy has, effectively, come true.

In the narrow sense, computers have not "passed the Turing Test." There is an annual competition, the Loebner Prize, that takes Turing's Imitation Game suggestion literally, pitting a panel of judges against chat programs and hidden human typists. It has been widely criticized for encouraging programs that use tricks such as simulated typing errors to fool the judges, instead of advancing the goal of making more intelligent machines. Even so, while several have fooled judges during extended conversation, none has yet won the prize.

More significantly, we now interact with artificial entities in daily life, often without realizing they are not human. In 1950, when Turing proposed the

Imitation Game, it was a stretch to think up a plausible scenario in which people would communicate via text with strangers of unknown and possibly fictitious identity. With the advent of the internet, this scenario has become commonplace.

In the mid-1990s, someone named Serdar Argic started inflaming the already heated Usenet arguments about the Armenian genocide by relentlessly posting hateful rants accusing the Armenians of massacring Turks. People wrote impassioned rebuttals to his screeds, thus making them even more disruptive by sidetracking any constructive discussion. Only after much anger and confusion did people realize that Argic was not a real person, but a program designed to intervene in any discussion that mentioned Armenia or Turkey, including Thanksgiving recipe posts. This was one of the first bots to deliberately fool people in a public setting.[16]

Chat-bots have since then become cleverer—and ubiquitous. They are tireless customer service agents, answering questions about ingredients, store hours, and mysterious error codes at any time of day or night. They are participants in online games, appearing as opponents, teammates, and incidental characters. They are the beautiful eager women in online dating sites who are always up for trying new things. Some are upfront about being software entities, but many attempt to pass as human.

An estimated 10–15 percent of users on the popular and influential social media site Twitter are bots. Some are useful: openly nonhuman programs that disseminate news, jokes, alerts, etc. But others masquerade as human users, seldom benevolently. They may be followers for hire, inflating their clients' apparent popularity. They may post vacation shots from sponsored villas, name-dropping restaurants, snacks, and songs, programmed to incessantly instigate flashes of envy and desire. Or they may be powerful purveyors of propaganda, chiming into political discussions, tirelessly hawking talking points, slogans, and manufactured rumors. Bots thrive here in part because Twitter limits posts to 140 characters;

---

[16] Judith Donath, *The Social Machine: Designs for Living Online* (Cambridge, MA: MIT Press, 2014).

non sequiturs, rather than back-and-forth discussions, characterize many interactions. Devising a program to mimic this style is much easier than creating one that must carry out an extended and coherent conversation.

Not all of today's artificial entities are online: we are increasingly surrounded by a growing population of social robots--autonomous, sentient-seeming objects. At home, we chat with friendly devices that fetch us the news, order us dinner, and ask politely about our day. We may have a robotic pet or coworker. There are robot receptionists who welcome guests in tech-forward hotels and robot orderlies who glide quietly into hospital rooms. Social robots are marketed as "friends" and "your next family member" who "can't wait to meet you."

No contemporary or readily foreseeable artificial entity is actually conscious or even primitively sentient, but our intuitive response to them is the opposite. They seem very much alert and aware. Our tendency to anthropomorphize contributes to this illusion. Yet when we see volition and intent in inanimate objects such as cars, trees or dolls, we recognize that we ourselves are the source of its imagined vitality. With artificial entities, the object itself behaves in ways that strongly suggest a sentient experience lies within.

The ambiguity of their identity—machine or new form of thinking being—is no accident. Like the chat-bots that score highly in the Loebner Prize competition by making spelling mistakes, social robots are often made to mimic human habits such as pausing or looking away as if thinking; these easy-to-implement tricks provide a convincing illusion of sentience. Many are designed with simple, round childlike curves--features that elicit nurturance, indulgence, and trust[17], while also keeping our expectations of their abilities low. Their gendered voices and linguistic insinuation of self-conscious thought ("I'd like to help you") give the impression that one is speaking to an aware and sentient

---

[17] Zebrowitz, Leslie. *Reading Faces*. (Boulder, CO: Westview Press, 1997).

being.[18] As Turing predicted, our use of language has changed: we casually speak of these entities wanting, thinking, and liking.

Ethics of Our Relationship with the Seemingly Sentient

What are the ethical issues involved in our interaction with artificial entities? One set of issues concerns our responsibilities toward them—how we should treat them. The ethical framework I will use here is based on Peter Singer's utilitarian applied ethics;[19] his sentience-focused approach to assessing responsibilities toward nonhumans make it especially relevant for thinking about artificial entities.[20] The key question here, however, is not how our treatment affects them, but what it does to us.

We noted earlier that our ethical responsibilities are to sentient beings: if something or someone has the capacity to feel, we need to take their preferences into consideration. To things that are not sentient—rocks, bacteria, dolls, robots—we have no direct moral obligation, that is, none that arises from their individual standing as a being with moral claims or rights. Since they do not experience anything, they cannot feel harmed by any action.

Though we do not have direct moral obligations to nonsentient entities, that does not mean we have no obligations toward them. Nonconscious entities have what are called "indirect rights." These are rights that come from their relationship to a being that does have ethical standing; because harming the nonconscious entity would harm the being with ethical standing, it should therefore should be

---

[18] Eyssel, Friederike, Laura De Ruiter, Dieta Kuchenbrandt, Simon Bobinger, and Frank Hegel. "'If You Sound Like Me, You Must Be More Human': On the Interplay of Robot and User Features on Human-Robot Acceptance and Anthropomorphism." Paper presented at the 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2012. 125-126

[19] Singer, Peter. *Practical Ethics* (Cambridge, UK: Cambridge University Press, 2011).

[20] The focus of this chapter is on Western society. See Kaplan, Frédéric. "Who Is Afraid of the Humanoid? Investigating Cultural Differences in the Acceptance of Robots." *International journal of humanoid robotics* 1, no. 03 (2004): 465-480 and Robertson, Jennifer. *Robo Sapiens Japanicus: Robots, Gender, Family, and the Japanese Nation*. Univ of California Press, 2017,, for reactions to artificial entities in Japan.

avoided. You adore your robot, and so I must treat it well because of your affection for it. It is wrong for me to harm something you value, not because of the intrinsic hurt to a thing (it has no feelings) but because you would be saddened by its loss.

Laws reflect a society's ethics, but they change slowly and are often more an indicator of the morals of its past. Indirect rights have been the primary source of protection that animals have had under American law: I cannot kick your dog, not because it would hurt your dog but because you would be upset (and it is your property). Indirect rights are often weak. In the moral calculus required to balance numerous competing preferences and rights, they can be readily eclipsed. Protection based on human preference disappears in the face of competing human interests—thus we have factory farms, sport hunting, etc.

Society changes. Laws protecting animals based on ethical reasoning that takes their experience into account—that recognizes their sentience—are becoming more common. The change is due both to (a) seeing sentience as the quality that defines whether one has direct moral claims and (b) recognizing that some animals are sentient. It is also part of a broader Western cultural shift to an increasingly inclusive view of who is a being with moral standing: it is not that long ago in the United States that women and slaves had mainly indirect rights. Advocates for animal rights posit that what they call "speciesism"—the belief that members of one species have superior moral standing on the basis of that membership—as the logical and moral equivalent of racism.

Some legal scholars have argued that such legal protection should extend to social robots:[21] "We may not want to be the kind of society that tolerates cruelty to an entity we think of as quasi-human."[22] I argue that this movement toward more inclusive rights does not, and should not, apply to nonsentient artificial beings. The fundamental reason for extending moral rights to animals is

---

[21] Darling, Kate. "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects." In *Robot Law*, edited by A. M. Froomkin R. Calo, & I. Kerr. (Cheltenhem: Edward Elgar Publishing, 2016) 213-234.

[22] Calo, Ryan. "Robotics and the Lessons of Cyberlaw." *California Law Review* (2015): 513-563.

recognition of their sentience—that they can experience suffering. It is a right inherent to them, regardless of whether a human observer, owner, or other interested party is aware of their pain.[23] The premise that sentience is the foundation of moral rights is important—extending these rights to nonsentient entities dilutes its meaning and significance.

That said, the compelling simulation of sentience exhibited by artificial entities can provide them with additional indirect moral claims, again stemming from considerations about a person's experience, not the entity's. Here the concern is that treating another cruelly brutalizes oneself. This principle is reflected in Jewish custom, which forbids sport hunting because it encourages cruelty, even if the animal is killed painlessly.[24] And Immanuel Kant, though he argued that animals have no "will" and thus no inherent rights, also wrote, "If he is not to stifle his own feelings, he must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men."[25]

Behaving ethically often involves trade-offs between competing rights and principles, and even a seemingly simple injunction such as "do not treat sentient-seeming entities cruelly" can create dilemmas. The popular keychain pet toy, the Tamagotchi, provides a useful scenario. These are very simple artificial entities that nonetheless exert a powerful emotional pull.[26] The owner of a Tamagotchi must work at keeping it "alive," a task that entails pushing buttons on it at frequent but arbitrary times. Ignore it and it will cease to thrive and will

---

[23] Darling points out that animal protection law seems to reflect the popular sentimental standing of particular animals, rather than the philosophically or biologically based concern with their sentience. In this chapter, our focus is on fundamental ethics—on getting the theory right in order to guide the practice.

[24] Rabbi Dr. Asher Meir, "Judaism and Hunting," *Jewish Ethicist*, https://www.ou.org/torah/machshava/jewish-ethicist/judaism_and_hunting/.

[25] Kant, Immanuel, and Louis Infield. *Lectures on Ethics*. Translated by Louis Infield. (New York: Harper & Row, 1963). 240.

[26] Kaplan, Frédéric. "Free Creatures: The Role of Uselessness in the Design of Artificial Pets." Presented at the *1st edutainment robotics workshop*. (Sankt Augustin, Germany:, 2000), 45-47.

eventually "die"; as with real pets, cruelty toward the Tamagotchi can take the form of neglect. Imagine now a family dinner. The grandmother is visiting, but a grandchild is continuously distracted, checking a Tamagotchi's status. Should the parents demand the child put the toy away and pay full attention to the (living, conscious, and closely related) grandparent present in the room, who would like their attention, but at the cost of allowing the Tamagotchi to possibly die? Or is nurturing the keychain pet useful training in responsible caring, so grandmother and virtual pet will need to share the child's divided attention?

The appeal of the simple Tamagotchi vividly demonstrates just how compelling and potentially manipulative an artificial entity can be. This raises concerns about prohibitions against mistreating them—and especially about encasing such prohibitions in law. The makers of an artificial entity can design it so that arbitrary events and conditions cause it to express suffering. The Tamagotchi appears to suffer because no one pressed its button at the demanded time. A more venal entity could appear to suffer when you do not purchase the items it is selling on behalf of the company that controls it; perhaps it will suffer unless it is taken on a Caribbean vacation, or it will appear to be lonely and unhappy if it is not in a room with you, recording all your conversations. The concept that we should not mistreat even a nonsentient entity because of the harm it does to ourselves is sound—but we need to be careful about who defines what is "cruel" in the arbitrary realm of artificial entities.

We should treat artificial entities at a minimum without cruelty—that is, without inflicting unnecessary harm to them. But what sort of relationship do we want to have with them? Here our concern shifts from sentience to consciousness.

"Your Next Friend Could Be a Robot" was the headline of a 2016 *Wall Street Journal* article that lauded the ease with which people become emotionally attached to social robots, a tendency that it claimed could solve, or at least ameliorate, the problem of loneliness among the elderly and the childless. The robots, the article notes, are far from intelligent, but they are "enhanced by the right auditory and visual cues" to seem like, as one social robot product manager said, "[a] likable person people want to have in their homes."

Such cues work, at least for the many people who express considerable affection for their social robots. A customer review for Alexa, Amazon's virtual assistant, says, "I wake up in the morning and she does the routine I've set up, and she's so comforting and useful and fun overall…feels like a new little buddy in the home."[27] A veteran technology writer described his relationship with social robot Jibo: "I work from home, and it's nice to have someone ask me how I'm doing when I'm making lunch." When the company behind it went out of business, his wrote of his heartbreak at its pending demise: "I've felt crushed knowing that every word the robot says to me could be his last," a heartbreak he compared with the loss he felt when his mother died after suffering from dementia.[28]

Though still a nascent technology, it is clear that people enjoy interacting with social robots. In coming years, we will have a growing number of relationships with artificial pets, coworkers, caretakers, and companions—and those bonds will become tighter as advances in machine learning, aided by the vast databases of user behavior metrics that existing entities have been able to collect, will make interacting with them ever more seamlessly polished and highly personalized.[29]

Not everyone sees this as a positive development. Technology and society researcher Sherry Turkle has written extensively about the ethical hazards of accepting artificial creations as personal companions, asking, "What is the value of interactions that contain no understanding of us and that contribute nothing to a shared store of human meaning?"[30] She has warned that robot companions may provide such a pleasant imitation of human company, without the inevitable disagreements and irritations that come with real people, that we may come to

---

[27] https://www.amazon.com/gp/customer-reviews/R2SSM75HH2PJD6/.

[28] Jeffrey van Camp, "My Jibo Is Dying and It's Breaking My Heart," *Wired*, March 8, 2019.

[29] Donath, Judith. "The Robot Dog Fetches for Whom?". In *A Networked Self and Human Augmentics, Artificial Intelligence, Sentience*, edited by Zizi Papacharissi. (London: Routledge, 2018) 26-40.

[30] Sherry Turkle, "Authenticity in the Age of Digital Companions," *Interaction Studies* 8, no. 3 (2007) 10-24.

prefer their frictionless companionship, whether as babysitters, friends, sexual partners, or caregivers, to that of a real, imperfect human being.

It was a virtual therapist—ELIZA—that pioneered relating socially with a machine, and it was people's enthusiastic reception of this same virtual therapist that prompted, in ELIZA's creator, the first backlash against such technologies. And virtual therapy provides a useful lens for examining the broader question of the values and ethics of forming a relationship with an artificial entity.

Although ELIZA was modeled after a "Rogerian psychiatrist," a computer therapist is antithetical to Carl Rogers's theory of psychology. In a 1977 profile, science writer Constance Holden outlined Rogers's main tenets: the therapist must be empathic (have "the ability to get inside the world of the client" and "see things as they look to him"), authentic (must "relate to the client as a person" and "allow himself to become involved with his feelings as well as his intellect"), and nonjudgmental ("let the client know he is accepted").[31] These guidelines address not how therapists should act but how they should think and feel; that they are capable of doing so is implicit.

Holden accompanied the profile of Rogers with a sidebar about ELIZA, titled "The Empathic Computer," which she concluded by noting, "Many lessons could be drawn from this, one of which is that even the appearance of empathy (combined, of course, with the computer's quite genuine nonjudgmentalism) can be extraordinarily powerful."[32] Weizenbaum sharply disagreed. Responding to this article, he quoted Rogers's argument that to effect a cure, the therapist must genuinely like the patient. "Of what help," he asked, "could it possibly be to anyone to know that he is worthy of being liked *by a computer*?" Weizenbaum concluded by saying: "The power of which Holden writes in connection with my

---

[31] Constance Holden, "Carl Rogers: Giving People Permission to Be Themselves," *Science* 198, no. 4312 (1977): 31-35.

[32] Constance Holden, "The Empathic Computer," *Science* 198, no. 4312 (1977): 32.

computer program is no more and no less than the power to deceive. No humane therapy of any kind ought to be grounded on that."[33]

Today, thousands of people confide their problems to virtual therapists. Some of the reasons are practical. The U.S. Defense Department, faced with thousands of veterans returning home suffering from PTSD and other psychological injuries, has supported the development of artificial therapists to relieve the acute shortages of human ones. Virtual therapy is far cheaper and more convenient, accessible wherever and whenever you need it.

Though the technology is still exploratory, studies indicate that therapy with an artificial entity is not only cost-effective but psychologically effective—and well liked.[34] In particular, people liked that the computer therapist was nonjudgmental: they were willing to divulge more personal information to it and to talk more freely about uncomfortable subjects, an openness that is invaluable in therapy.[35]

If openness and honesty are the desired behaviors in therapy, then why have a therapist at all? Why not just have a pure text interface, with no artificial therapist, no implied yet nonexistent being? The answer is that the personified interface, with its imagined therapist, is more engaging; it inspires people to interact with it more and to attend to its suggestions. For example, *Woebot* is a conversational entity that provides cognitive-behavioral therapy via text chat; it has been found to significantly reduce depression in its users, who say they like its personality, and that it pays attention to them and holds them accountable for

---

[33] Joseph Weizenbaum, "Computers as 'Therapists'," *Science* 198, no. 4315 (1977): 54.

[34] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile, "Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial," *JMIR Ment. Health* 4, no. 2 (2017): e19; Gale M. Lucas et al., "It's Only a Computer: Virtual Humans Increase Willingness to Disclose," *Computers in Human Behavior* 37 (2014): 94-100; Adam S. Miner, Arnold Milstein, and Jefferey T. Hancock, "Talking to Machines about Personal Mental Health Problems," *JAMA* 318, no. 13 (2017): 1217-1218.

[35] Lucas et al., "It's Only a Computer."

being attentive to their emotions.[36] Though its interface is quite simple, the user's mental model of engaging with an entity provides a quite different experience than would a similar interaction framed as an interactive questionnaire. The ersatz empathy that Weizenbaum decried turns out to be valuable after all.

We humans are highly social beings, and in the presence of others—even imagined others-- we try, for better or worse, to make a desired impression. Studies comparing how people respond to questions asked by a computer with a facial versus a text interface found that they are more responsive and engaged with the facial interface--but also less honest, painting themselves in a more favorable light.[37] Hints of personhood, of approval or displeasure, influence how we act.

For understanding our relationship with artificial entities in general, the most significant observation is that the virtual therapist plays a novel role, one that could be played neither by a human nor by a simple questionnaire. People are aware that the virtual therapist is artificial and not conscious, so they feel comfortable confiding in it, yet they can at the same time suspend this recognition and engage with it as if it were a conscious and empathic being. Designing the ideal virtual therapist means balancing being engaging (more humanlike) against inviting candid disclosures (more machinelike) to create an exemplar not found in nature.

Yet the relationship between therapist and patient is a particular kind of relationship, and we want to be careful about the parallels we draw to friendships and other social bonds. It is possible, at least in some forms of therapy, to cast the therapeutic relationship as instrumental, even commercial: the patient pays the therapist to perform the service of helping them with their mental health; the relationship is a success if the patient's health improves. (This is, of course, deeply at odds with how Rogers, Weizenbaum, and many others understand the

---

[36] Fitzpatrick, Darcy, and Vierhile, "Delivering Cognitive Behavior Therapy to Young Adults."

[37] L. Sproull et al., "When the Interface Is a Face," *Human Computer Interaction* 11 (1996): 97-124.

therapeutic process.)[38] The point is that when a relationship is seen as purely or primarily instrumental—with the party receiving the service unconcerned with the thoughts of the one performing it and interested only in the outcome—then substituting an artificial entity into the role of service provider makes sense.[39] This is especially so when, as is the case with patients uncomfortable with the possibility of being judged and looked down upon by a therapist, being thought about by the service provider is seen as negative.

Our relationships are a mix of nurturing bonds and instrumental uses in varying proportions. Nurturing holds society together, and it is fundamental to who we are as humans. We evolved to nurture, to derive joy from taking care of others and knowing that we have made them happy: we take care of our family, our friends, our pets, and our plants.

Yet, for a variety of reasons—an emphasis on efficiency, the anonymity of city life, an industrialized corporate service economy—we now live in a world where many formerly social and engaged relationships are recast as instrumental ones[40], transformed from ones where a robot would be a poor substitute to ones where there is little care or empathy left to lose.

We need to be cognizant of the sometimes subtle but fundamentally important empathic and bonding element of our relationships, to care not only about what the relationship can do for us but also about how we affect the other—to care about both the experience of the other and the other's thoughts of us. It is possible to measure the usefulness of these bonds, to quantify the health or productivity increase they provide, but that is only a piece of their value.

---

[38] Cecil Holden Patterson, "Empathy, Warmth, and Genuineness in Psychotherapy: A Review of Reviews," *Psychotherapy: Theory, Research, Practice, Training* 21, no. 4 (1984): 431-438.

[39] We are omitting here the quite significant ethical issue of robot-induced unemployment. John Danaher, "Will Life be Worth Living in a World without Work? Technological Unemployment and the Meaning of Life," *Science and Engineering Ethics* 23, no. 1 (2017): 41-64.

[40] Hochschild, Arlie Russell. *The Managed Heart: Commercialization of Human Feeling*. (Berkeley, CA: Univ of California Press, 2012). 1983; Donath, Judith. "Our Evolving Super-Networks." In *The Social Machine: Designs for Living Online*. (Cambridge, MA: MIT Press, 2014) 111-132.

That other-centric element is absent in interactions with an artificial entity, leaving only the instrumental element—how does this relationship benefit me? Such entities, and thus such relationships, will play an increased role in our lives in the coming years. Weizenbaum's fears about our willingness to embrace machines was prescient—it is perhaps ironic that virtual therapy may be the one applications in which the machine's absence of mind is truly beneficial.

Ethics of Creating Seemingly Sentient Entities

We have been focusing thus far on the ethics of our relationships with artificial entities. We turn now to the process creating these entities, and in particular, of designing them to seem conscious and aware when they are not. Here, the ethical questions center on deception.

There is an extensive, and contentious, body of work about the ethics of deception.[41] The central questions are: What exactly constitutes a deception? Are all deceptions ethically wrong—and if not, which ones are permitted and why? For the purpose of this discussion, I will put forth some basic definitions and ethical premises, so we can focus on the new issues artificial entities raise.

An act or quality is deceptive if it is intended to cause the recipient to believe something that is not true. Intent is key: not every false statement or causing of false belief is deceptive. If one believes something that is not true, and tells that untrue thing to others, that is a mistake, not a deception. If one says something true, but the recipient misconstrues or misinterprets it, that is a misunderstanding, not a deception.

Ethical concerns focus on intentional deceptions. While a mantis that evolved to resemble a dead leaf is deceptive and this deception harms its predators,

---

[41] See, e.g., Sissela Bok, *Lying: Moral Choice in Public and Private Life* (New York: Vintage, 1999); Bella M. DePaulo et al., "Lying in Everyday Life,"*Journal of Personality and Social Psychology* 70, no. 5 (1996): 979–995; D. B. Buller and J. K. Burgoon, "Interpersonal Deception Theory," *Communication Theory* 6 (1996): 203–242; Hancock, Jeffrey T. "Digital Deception." In *Oxford Handbook of Internet Psychology*, edited by Katelyn McKenna, Adam Joinson, Tom Postmes, Ulf-Dietrich Reips. (Oxford, UK: Oxford University Press, 2007) 289-301.

it is not unethical, for the mantis did not choose to deceive. Humans lie deliberately--and so do some animals; it is a sign of advanced cognition.

A few philosophers have declared all lying to be immoral. St Augustine declared all lies to be sinful; Kant said, "To be truthful (honest) in all declarations, therefore, is a sacred and absolutely commanding decree of reason, limited by no expediency"; and Sam Harris a a contemporary proponent of radical honesty, challenges his readers to abstain from any and all lies.[42]

Most people (and philosophers) hold more nuanced, if differing, views, evaluating the ethics of deceptions by the harm they cause. An altruistic deception is done at one's own expense to benefit the other; a selfish deception is done for one's own gain and harming the recipient is an effect but not the goal; a malicious deception is performed with the goal of harming the recipient. In an ethical calculus of deception, one might argue that altruistic deceptions are ethical, and ones that cause harm should be assessed based on the amount of harm caused and the moral standing of the various parties. A lie to a would-be mass shooter that results in his capture and saves many lives is by narrow definition a malicious lie, but most people would agree that it was ethical.

Many of the issues concerning deception and artificial entities are analogous to or instances broader ethical controversies. For example, Paro is an artificial baby harp seal: cuddly, responsive, and lifelike. Is it ethical to give Paro to elderly dementia patients, who believe it is really alive?[43] This can should be considered in the context of the larger ongoing debate about the ethics of deceiving such

---

[42] Hermanowicz, Erika T. "Augustine on Lying." Speculum 93, no. 3 (2018/07/01 2018): 699-727; Kant, Immanuel. "On a Supposed Right to Lie from Altruistic Motives," Translated by Lewis White Beck, in *Immanuel Kant: Critique of Practical Reason and Other Writings in Moral Philosophy*, translated and edited by Lewis White Beck. (Chicago: University of Chicago Press, 1949) 346-350; Sam Harris, *Lying* (US: Four Elephants Press, 2013).

[43] Shannon Vallor, "Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-First Century," *Philosophy & Technology* 24, no. 3 (2011), 251-268; Angela Johnston, "Robotic Seals Comfort Dementia Patients but Raise Ethical Concerns," in *Crosscurrents* (San Francisco, CA: KALW, 2015).

patients with the goal of calming and reassuring them[44]. If one concludes that any deception that provides comfort to such patients is permissible, that would apply to Paro, too.[45]

Identity deception of some kind is inherent to all artificial, seemingly sentient entities: they are made to look, act, and/or speak as if a thinking, feeling, sensing mind was motivating them. Even for one to declare "I am a program" is, arguably, deceptive, for the use of the word "I" implies a thinking self-aware existence, the being whose thought process formed those words.[46] Note that the responsibility for the deception lies with the person who initiated it, not the medium that conveyed it; the artificial entity is no more responsible for its deceptions than is a note saying, "The dog ate my homework."

The identity presentation of artificial entities spans a range from fairly transparent to fully deceptive. Physical robots are, thus far, clearly artificial. Though they may have features such as a human-like voice, eyes that follows us

---

[44] E.g. MacFarquhar, Larissa. "The Comforting Fictions of Dementia Care." *New Yorker* (2018): 42-55.

[45] Another ethical issue about Paro and other "carebots," is concern about offloading caregiving to a machine. Technologies that *assist* human caregivers may be greatly beneficial to all, but using them to *replace* human care harms not only the patient but also, as Vallor argues, the caregivers (Vallor, "Carebots and Caregivers"). We need to be careful not to think of caregiving as only a burdensome task but also to keep in mind the importance of nurturing as a human and humane quality. See, more generally, Arlie Russell Hochschild, *The Outsourced Self: Intimate Life in Market Times* (New York: Metropolitan Books, 2012).

[46] There is a worldview in which artificial entities are arguably not deceptive. Sociologist Erving Goffman posited that society functioned much like theater: we play roles, with greater or lesser skill, adapting them to different situations. In this theater of everyday life, we act in public in ways that are at odds with how we feel, saying the polite thing even when it is not true, wearing the clothes and voicing the opinions the role we are playing demands. Acting is not deception, because the audience does not permanently believe it—they "suspend" (real) belief; rather, this role-playing is beneficial, even necessary, because it enables us to live together more or less harmoniously. Goffman, Erving. "On Face-Work: An Analysis of Ritual Elements in Social Interaction," in *Interaction Ritual* (New York: Pantheon Books, 1967), previously published *Psychiatry: Journal of Interpersonal Relations*, 18, no 3 (1955): 213-231; *The Presentation of Self in Everyday Life* (Garden City, NY: Doubleday Anchor Books, 1959). One might argue artificial entities are performing sentience, but we understand this to be a role, much as everyone is playing, and not a deception.

across the room, little gestures, etc. that lead us—or deceive us—to think of them as individuals with distinct personalities, we do not mistake them for humans or animals.[47] Online, however, software agents easily pass as human in contexts where conversations come in short and sometimes cryptic bursts. Where there is no tell-tale physical body, the possibility for deception is much higher.

An entity that disseminates dangerous propaganda or other information with malicious intent is easy to classify as unethical, regardless of whether it deceptively claims to be human or honestly declares itself a bot (though the former is likely to be more persuasive and thus more harmful.)[48]

A harder question concerns the ethics of identity deception performed for benevolent purposes. Is it ethical to create, say, a bot that patrols discussion sites correcting erroneous medical information while masquerading as a doctor to establish its authority? An absolutist would declare this, like any other deception, unethical. At the other extreme, a utilitarian might argue that because the identity deception has beneficial effects, such a falsehood is permissible—perhaps even required. While impersonating a doctor, even with good intentions, is usually judged to be unethical, one reason is that we assume that the impersonator is not qualified to provide the advice and is making a false identity claim in order to be accorded trust which they do not deserve. While that is likely when dealing with human impersonators, it may not apply to a bot--what if its medical knowledge is greater than any human's?

---

[47] The easy recognition of robots may be temporary: several research labs work on creating robots that look as humanlike as possible, e.g. Ishiguro, Hiroshi, and Shuichi Nishio. "Building Artificial Humans to Understand Humans," in *Geminoid Studies: Science and Technologies for Humanlike Teleoperated Androids*, edited by Hiroshi Ishiguro and Fabio Dalla Libera. (Singapore: Springer Nature, 2018) 21-37; Hanson, David. "Exploring the Aesthetic Range for Humanoid Robots." Paper presented at the Proceedings of the ICCS/CogSci-2006, Vancouver, British Columbia, July 26-29 2006. And Paro looks remarkably like a baby seal, though its behavior is certainly different.

[48] Ishowo-Oloko, Fatimah, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. "Behavioural Evidence for a Transparency–Efficiency Tradeoff in Human–Machine Cooperation." *Nature Machine Intelligence* 1, no. 11 (2019/11/01 2019): 517-521.

In considering whether "beneficial" deceptions are ethical, the notion of *autonomy* is central[49]. It is a concept most familiar from debates about patient/doctor communication.[50] For many years, Western doctors followed a practice of paternalistic utilitarianism, assuming that persuading the patient to comply with their treatment recommendations was ethical regardless of the means, including withholding information or lying to patients about their condition. More recently, patients and some philosophers have challenged this view, arguing that patients have the right to autonomy-- to make informed decisions for themselves.

Artificial entities generate analogous dilemmas. If people would follow the advice of a respected person, but not a bot, is it ever ethical to make the bot mimic that person (or type of person) in order to gain credibility, even for a good cause? We mentioned above that in the utilitarian view, such mimicry could be seen as beneficial; the principle of autonomy, however, says that taking away someone's ability to make their own unmanipulated judgements is an ethical violation in itself.

When we receive information from others, whether it be news of the world, advice, local gossip, etc. our assessment of its veracity is often based on whether we trust its source: do we believe they are knowledgeable and that they do not have ulterior motives to harm us? Identity deception manipulates that trust, inducing us to believe things we otherwise would not.

What does it mean to trust an artificial entity? It is easy to slip into thinking of the artificial entities themselves as deceptive or trustworthy, but they are a medium, not a mind—a conduit for the goals of human designers, owners and controllers. When we meet people, we try to figure out their identity—their role in society—in order to make sense of who they are, what motivates them and what they may be seeking in the interaction. The analogous questions regarding a

---

49 Bok, Sissela. *Lying: Moral Choice in Public and Private Life*. (New York: Vintage, 1999).

50 Sokol, Daniel K. "Can Deceiving Patients Be Morally Acceptable?". *The BMJ* 334, no. 7601 (2007): 984-86.

robot are not "What does it want?" but "Who controls it?" and "Who has access to the data it collects and what is *their* motivation?"

Today, very few artificial entities are self-contained; most exist in frequent dialog with a larger, more powerful system, which may assist the with interpreting speech, analyzing images, or other computation-heavy tasks. Not all have remote "brains": a Tamagotchi, for example, is a self-contained toy and I can run an instance of ELIZA on my own computer, and our conversations will be private between us. But many artificial entities have their real brain at—or at least send their data to—a distant location.

This introduces privacy-related ethical questions. If I confide in an artificial therapist because I am more comfortable discussing my problems with a machine, I may be quite discomforted to find out that my words are in fact uploaded, read, and analyzed by people.[51] If I type a search query into Google, I understand that the query goes to some distant computer; but if I ask a question of the companionable entity sitting on my kitchen counter, my sense is that the creature is answering, not that it is sending that query to some distant location—though that is indeed what is happening. The design of artificial entities encourages us to think of them as independent beings, not, as most of them are, front-end interfaces to an extensive computer system.

Some artificial entities gather extensive data about their users, recording conversations, eye movements, and gestures; ensconced in a living space, they can collect contextual information about how the people in their purview respond to a wide range of events. If this data is collected only to improve interactions with the person—say, to understand their accent better—one may judge it useful and acceptable. But the goals of the robot—or more accurately the robot's controllers' goals—may diverge sharply from the goals of the user. The entertaining toy or trusted companion's ulterior purpose may be to sell goods, promote a viewpoint, or otherwise to influence one's opinions, wants, and behavior. And such entities may become extraordinarily effective persuaders.

---

[51] Lucas et al., "It's Only a Computer."

An active and growing field of research seeks to understand how to design technologies that influence people and compel them to conform and obey. Robots that "use human-like gazing behavior" are known to be persuasive—and become even more so if gestures are added.[52] If a robot does something that induces gratitude, "the norm of reciprocity compels people to return a favor."[53] People conform when faced with "active peer pressure" from a group of robots,[54] and "robots have enough authority to pressure participants, even if they protest, to continue a tedious task for a substantial amount of time."[55] The published research cites laudable goals as potential applications; the technology will help the user stick to a diet, follow crucial directions, or use environmentally responsible products. Yet there is nothing that ensures that these powerful techniques will always be used so benevolently.

Sentient Entities as Social Mirror

The big-eyed, round-bodied artificial assistant that sits on our counter, playing music and telling jokes, seems disarmingly innocuous; if we think of it as having intentions, they are to please us. But when we imagine an actually sentient, conscious artificial being and its goals and intentions, the narrative tends to darken. To understand why, we need to turn to another mental quality—intelligence.

Vernacular Western thought pictures the world as hierarchical, with humans on top due to our superior intelligence. This intelligence has given us fantastic

---

[52] Jaap Ham et al., "Making Robots Persuasive: The Influence of Combining Persuasive Strategies (Gazing and Gestures) by a Storytelling Robot on Its Persuasive Power," paper presented at the Social Robotics: Third International Conference, ICSR, Amsterdam, The Netherlands, November 24–25, 2011.

[53] Lee, Seungcheol Austin, and Yuhua Liang. "The Role of Reciprocity in Verbally Persuasive Robots." *Cyberpsychology, Behavior, and Social Networking* 19, no. 8 (2016): 524-527.

[54] Katsila, Athanasia. "Active Peer Pressure in Human-Robot Interaction." Masters thesis, University of Nevada 2018.

[55] Denise Y Geiskkovitch et al., "Please Continue, We Need More Data: An Exploration of Obedience to Robots,"*Journal of Human-Robot Interaction* 5, no. 1 (2016), 82-99.

power: we build bridges, cities, bombs, and transistors; we conquer nature with vaccines, dams, and insecticides. Our intelligence has given us power over all the other animals, which we have exploited without hesitation. But while our intelligence gives us the ability to do these things, it is our consciousness—our awareness of ourselves, our place in the world, and our future—that has provided the ambition to do so.

So long as the machine is merely intelligent, cleverly solving very difficult problems—problems far too complex for a mere human intelligence to solve—it does not pose an existential threat to us. It solves the problems simply because that's what it automatically, mindlessly does, much like a bacterium reverses course away from an obstacle. But if that machine somehow becomes sentient, with preferences and the drive to achieve them—or conscious, with a sense of self and of the future, the ingredients for ambition—then it is deeply threatening to us.

In our imagination, at least as shaped by our modern, capitalist, Western way of thinking, that ambition must inevitably be to dominate—to be the alpha, the top of the food chain. We achieved this pinnacle with our superior intelligence—and a super-intelligent machine, far smarter than ourselves, will, we assume, use that intelligence to supersede us.

Samuel Butler voiced this fear in his 1872 novel *Erewhon*: "The machines were ultimately destined to supplant the race of man, and to become instinct with a vitality as different from, and superior to, that of animals, as animal to vegetable life." To prevent this destiny, the people of Erewhon destroyed all machines and banned their manufacture[56]. Karel Capek introduced the word "robot" in *R.U.R,* his 1920 play in which the robots, provoked by long mistreatment, rise in rebellion and ultimately annihilate the human race. In the 1967 short story "I Have No Mouth, and I Must Scream," Harlan Ellison describes a world in which humanity has been made nearly extinct by intelligent machines that had been programmed to wage war; the few humans that remain are tormented by the sadistic and now conscious AIs.

---

[56] It is a satirical novel, and whether the world it presents is utopian or dystopian is ambiguous.

Today, the fear that the machines we make will supplant us echoes in warnings not only from science fiction writers and technology critics, but from scientists and engineers themselves. Physicist Stephen Hawking warned that "The development of full artificial intelligence could spell the end of the human race". Along similar lines, inventor Elon Musk has said "If AI has a goal and humanity just happens to be in the way, it will destroy humanity as a matter of course without even thinking about it."[57]

It is not certain that a machine can ever become sentient or conscious; even if it could, it is far from known by what process or—dystopian terrors aside—what sort of being it will be. All the conscious beings we know are living creatures, who evolved over millions of years through a process that favored reproductive survival. A machine consciousness would be vastly different, in ways we cannot predict.[58]

Consciousness, as we have discussed, is an enigmatic property. Unable to precisely measure or even to define it, our assessments of other beings' consciousness are heavily shaded by our preferences and conveniences. We erroneously ascribe emotions and an inner life to nonsentient but humanoid machines, while vastly underestimating the inner life of animals, denying their sense of self, even their ability to feel pain.[59] Motivating this willful ignorance is the immense profit that comes with asserting that all other creatures exist for

---

[57] Cellan-Jones, Rory. "Stephen Hawking Warns Artificial Intelligence Could End Mankind." *BBC news*, December 2, 2014; Browne, Ryan. "Elon Musk Warns A.I. Could Create an 'Immortal Dictator from Which We Can Never Escape'." *CNBC*, April 6, 2018.

[58] Much speculations about AI posit that consciousness would emerge out of sufficient complexity; see, e.g., M. Minsky, *The Society of Mind* (New York: Simon & Schuster, 1988), though if we look at the biological record it seems that basic sentience arose with pain and pleasure, as the ability to experience emotion in response to sensory input. For an emergent computational mind, the negative and positive inputs need not be imitations of the organic forms—perhaps its native valences would be the billions of likes and dislikes that are registered across the internet.

[59] Frans De Waal, *Are We Smart Enough to Know How Smart Animals Are?* (NewYork: WW Norton & Company, 2016); Steiner, Gary, *Anthropocentrism and Its Discontents: The Moral Status of Animals in the History of Western Philosophy* (Pittsburgh: University of Pittsburgh Press, 2010).

humans to use—to be made into food and clothing, to carry burdens, to test medicines, and to entertain us—and the relief from responsibility that comes with insisting, even in the face of vivid contrary evidence, that they are incapable of suffering.

Our dystopian predictions of what a powerful and conscious machine would do are not based on projection from the technology or even from biology. They seem, instead, like the nightmares of a guilty conscience. The ethical challenge is to use this existential guilt to change. Can we treat the other beings we live with on Earth as we would want conscious, super-powerful artificial entities to treat us?

Bibliography

Broom, Donald M. Sentience and Animal Welfare. CABI, 2014.

Calo, Ryan. "Robotics and the Lessons of Cyberlaw." California Law Review (2015): 513–63.

DePaulo, Bella M., Deborah A. Kashy, Susan E. Kirkendol, Melissa M. Wyer, and Jennifer A. Epstein. "Lying in Everyday Life." Journal of Personality and Social Psychology 70, no. 5 (1996): 979.

Donath, Judith. "The Robot Dog Fetches for Whom?" In A Networked Self and Human Augmentics, Artificial Intelligence, Sentience, 26–40: Routledge, 2018.

Godfrey-Smith, Peter. Other Minds: The Octopus and the Evolution of Intelligent Life. London: William Collins, 2016.

Kaplan, Frédéric. "Who Is Afraid of the Humanoid? Investigating Cultural Differences in the Acceptance of Robots." International Journal of Humanoid Robotics 1, no. 3 (2004): 465–80.

Singer, Peter. Practical Ethics. Cambridge: Cambridge University Press, 2011.

Turing, Alan. "Computing Machinery and Intelligence." Mind 49 (1950): 433–60.

Turkle, Sherry. "Authenticity in the Age of Digital Companions." Interaction Studies 8, no. 3 (2007): 501–17.

Weizenbaum, Joseph. Computer Power and Human Reason. San Francisco: W. H. Freeman, 1976.

Weizenbaum, Joseph. "Contextual Understanding by Computers." Communications of the ACM 10, no. 8 (1967): 474–80