10-3-2017

Artificial Intelligence and the Ethics of Self-learning Robots

Shannon Vallor
Santa Clara University, svallor@scu.edu

George A. Bekey

Recommended Citation

Vallor, S., & Bekey, G. A. (2017). Artificial Intelligence and the Ethics of Self-learning Robots. In P. Lin, K. Abney, & R. Jenkins (Eds.), Robot Ethics 2.0 (pp. 338–353). Oxford University Press.

22 ARTIFICIAL INTELLIGENCE AND THE ETHICS OF SELF-LEARNING ROBOTS

Shannon Vallor and George A. Bekey

The convergence of robotics technology with the science of artificial intelligence (or AI) is rapidly enabling the development of robots that emulate a wide range of intelligent human behaviors.[1] Recent advances in machine learning techniques have produced significant gains in the ability of artificial agents to perform or even excel in activities formerly thought to be the exclusive province of human intelligence, including abstract problem-solving, perceptual recognition, social interaction, and natural language use. These developments raise a host of new ethical concerns about the responsible design, manufacture, and use of robots enabled with artificial intelligence—particularly those equipped with self-learning capacities.

The potential public benefits of self-learning robots are immense. Driverless cars promise to vastly reduce human fatalities on the road while boosting transportation efficiency and reducing energy use. Robot medics with access to a virtual ocean of medical case data might one day be able to diagnose patients with far greater speed and reliability than even the best-trained human counterparts. Robots tasked with crowd control could predict the actions of a dangerous mob well before the signs are recognizable to law enforcement officers. Such applications, and many more that will emerge, have the potential to serve vital moral interests in protecting human life, health, and well-being.

Yet as this chapter will show, the ethical risks posed by AI-enabled robots are equally serious—especially since self-learning systems behave in ways that cannot always be anticipated or fully understood, even by their programmers. Some warn of a future where AI escapes our control, or even turns against humanity (Standage 2016); but other, far less cinematic dangers are much nearer to hand and are virtually certain to cause great harms if not promptly addressed by

technologists, lawmakers, and other stakeholders. The task of ensuring the ethical design, manufacture, use, and governance of AI-enabled robots and other artificial agents is thus as critically important as it is vast.

22.1 What Is Artificial Intelligence?

The nature of human intelligence has been one of the great mysteries since the earliest days of civilization. It has been attributed to God or civilization or accidental mutations, but there is general agreement that it is our brain and the intelligence it exhibits that separates humans from other animals. For centuries it was thought that a machine would never be able to emulate human thinking. Yet at present there are numerous computer programs that emulate some aspect of human intelligence, even if none can perform all the cognitive functions of a human brain. The earliest computer programs to exhibit some behavioral aspect of intelligence began to appear in the second half of the twentieth century.[2] The first meaningful test of a computer's approximation to human intelligence was proposed by Alan Turing (1950). He called it the "imitation game," more commonly known today as the "Turing Test."

The idea is the following: An investigator submits written queries to the computer, which replies in writing. The computer passes the test if, after a suitable time interval, the average investigator has no better than a 70% chance of correctly determining whether the responses come from a person or a computer. The general utility and significance of the Turing Test for AI research are widely contested (Moor 2003; Russell and Norvig 2010). Its focus on a system's appearance to users in a tightly controlled setting, rather than the cognitive architecture or internal operations of the system, may appear to bypass the basic point of the test: namely, to demonstrate a *cognitive* faculty. The test also excludes many other types of intelligent performance that do not involve conversational ability. Still, it is noteworthy that in an annual competition held since 1991 (the Loebner Prize), no system has passed an unrestricted version of the test—repeatedly defying predictions by many researchers (Turing included) that computers would display conversational intelligence by the twenty-first century (Moor 2003).

While there are many unresolved questions about what it would take for a machine to demonstrate possession of "real" intelligence of the general sort possessed by humans, the chief goal of most AI researchers is more modest: systems that can emulate, augment, or compete with the performance of intelligent humans in well-defined tasks.[3] In this sense, the pragmatic legacy of the Turing Test endures. This figurative, task-delimited definition of artificial intelligence is the one we shall employ in the rest of this chapter, unless otherwise stated. It is distinct from the far more ambitious notion of "strong" artificial intelligence

with the full range of cognitive capacities typically possessed by humans, including self-awareness. Most AI researchers characterize the latter achievement, often referred to as "artificial general intelligence" or AGI, as *at best* a long-term prospect—not an emerging reality.[4]

Artificial agents with specific forms of task intelligence, on the other hand, are already here among us. In many cases they not only compete with but handily *outperform* human agents, a trend that is projected to accelerate rapidly with ongoing advances in techniques of *machine learning*. Moreover, the implementation of task-specific AI systems in robotic systems is further expanding the range and variety of AI agents and the kinds of social roles they can occupy. Such trends are projected to yield significant gains in global productivity, knowledge production, and institutional efficiency (Kaplan 2015). Yet as we will see, they also carry profound social, economic, and *ethical* risks.

22.2 Artificial Intelligence and the Ethics of Machine Learning

Many ethical concerns about AI research and its robotic applications are associated with a rapidly emerging domain of computer science known as *machine learning*. As with learning in animals, machine learning is a developmental process in which repeated exposures of a system to an information-rich environment gradually produce, expand, enhance, or reinforce that system's behavioral and cognitive competence in that environment or relevantly similar ones. Learning produces changes in the state of the system that endure for some time, often through some mechanism of explicit or implicit memory formation.

One important approach to machine learning is modeled on networks in the central nervous system and is known as *neural network learning* or, more accurately, *artificial neural network (ANN) learning*. For simplicity, we omit the word *artificial* in the following discussion. A neural network consists of a set of *input nodes* representing various features of the source or input data and a set of output nodes representing the desired control actions. Between the input and output node layers are "hidden" layers of nodes that function to process the input data, for example, by extracting features that are especially relevant to the desired outputs. Connections between the nodes have numerical "weights" that can be modified with the help of a *learning algorithm*; the algorithm allows the network to be "trained" with each new input pattern until the network weights are adjusted in such a way that the relationship between input and output layers is optimized. Thus the network gradually "learns" from repeated "experience" (multiple training runs with input datasets) how to optimize the machine's "behavior" (outputs) for a given kind of task.

While machine learning can model cognitive architectures other than neural networks, interest in neural networks has grown in recent years with the addition of more hidden layers giving *depth* to such networks, as well as feedback or recurrent layers. The adjustment of the connection strengths in these more complex networks belongs to a loosely defined group of techniques known as *deep learning*. Among other applications of AI—especially those involving computer vision, natural language, or audio processing—the performance of self-driving "robotic cars" has been improved significantly by the use of deep learning techniques.

Machine learning techniques also vary in terms of the degree to which the learning is *supervised*, that is, the extent to which the training data is explicitly labeled by humans to tell the system which classifications it should learn to make (as opposed to letting the system construct its own classifications or groupings). While many other programming methods can be embedded in AI systems, including "top-down" rule-based controls ("If a right turn is planned, activate right turn signal 75 meters prior to turn"), real-world contingencies are often too numerous, ambiguous, or unpredictable to effectively manage without the aid of machine learning techniques.

For a self-driving car, the inputs will include real-time data about road conditions, illumination, speed, GPS location, and desired destination. The outputs will include the computed values of controlled variables, such as pressure on the accelerator (gas) pedal, steering commands (e.g., "Turn the steering wheel 30 degrees clockwise"), and so on. Hidden layers of nodes will be sensitive to a wide range of salient patterns that might be detected in the inputs (e.g., input patterns indicating a bicyclist on the right side of the roadway) in ways that shape the proper outputs ("Slow down slightly, edge to the left-center of the lane").

Before it is capable of driving safely in real-world conditions, however, the car's network must be "trained" by a learning algorithm to predict the appropriate machine outputs (driving behaviors) for a wide variety of inputs and goals. Learning takes place by adjustment of the gains or *weights* between the nodes of the network's input, hidden, and output layers. Initial training of a network in simulations is followed by controlled field tests, where the network is implemented and trained in a physical car. Once the proper connection strengths are determined by the training process, the input-output behavior of the network becomes an approximation to the behavior of the system being modeled: in our example, a well-driven car. While an artificial neural network's cognitive structure may bear little resemblance to the neural structure of a competent human driver, once it can reliably approximate the input-output behavior typical of such drivers, we may say the network has *learned* to drive.

Once the network is judged sufficiently competent and reliable in controlled tests, additional fine-tuning of its performance might then take place "in the wild," that is, in uncontrolled real-world conditions—as in the case of Tesla's autopilot

feature. Here we begin to confront important ethical questions emerging from AI's implementation in a robot or other system that can autonomously act and make irreversible changes in the physical world. Media outlets have widely covered the ethics of autonomous cars, especially the prospect of real-world "trolley problems" generated by the task of programming cars to make morally challenging trade-offs between the safety of its passengers, occupants of other cars, and pedestrians (Achenbach 2015). Yet "trolley problems" do not exhaust or even necessarily address the core ethical issues raised by artificially intelligent and autonomous robotic systems.[5] This chapter focuses on a range of ethical issues less commonly addressed in media coverage of AI and robotics.

First, consider that driverless cars are intended to make roads safer for humans, who are notoriously unsafe drivers. This goal has *prima facie* ethical merit, for who would deny that fewer car wrecks is a moral good? To accomplish it, however, one must train artificial networks to drive *better* than we do. Like humans, self-learning machines gain competence in part by learning from their mistakes. The most fertile grounds for driving mistakes are real-world roadways, populated by loose dogs, fallen trees, wandering deer, potholes, and drunk, texting, or sleepy drivers. But is it ethical to allow people on public roads to be unwitting test subjects for a driverless car's training runs?

Tesla's customers voluntarily sign up for this risk in exchange for the excitement and convenience of the latest driving technology, but pedestrians and other drivers who might be on the wrong end of an autopilot mistake have entered into no such contract. Is it ethical for a company to impose such risks on us, even if the risks are statistically small, without public discussion or legislative oversight? Should the public be compensated for such testing, since Tesla—a private company—profits handsomely if the tests result in a more commercially viable technology? Or should we accept that since the advancement of driverless technology is in the long-term public interest, we (or our children) will be compensated by Tesla with vastly safer roads five, ten, or twenty years from now?

Moreover, does ethics permit the unwitting sacrifice of those who might be endangered *today* by a machine learning its way in the world, as long as we can reasonably hope that many others will be saved by the same technology tomorrow? Here we see an implicit conflict emerging between different ethical theories; a utilitarian may well license such sacrifices in the interests of greater human happiness, while a Kantian would regard them as fundamentally immoral. Similar questions can be asked about other applications of machine learning. For example, should a future robot medic, well trained in simulations and controlled tests, be allowed to fine-tune its network in the field with real, injured victims of an earthquake or mass shooting, who might be further endangered by the robot's error? Does the prospect look more ethically justifiable if we reasonably believe this will increase the likelihood of one day having *extraordinary*

robot medics that can save many more lives than we currently can with only human medics?

Imagine that we decide the long-term public benefit does justify the risk. Even the most powerful and well-trained artificial networks are not wholly predictable. Statistically they may be competitive with or even superior to humans at a given task, but unforeseen outputs—sometimes quite odd ones—are a rare but virtually ineradicable possibility. Some are *emergent behaviors* produced by interactions in large, complex systems.[6] Others are simple failures of an otherwise reliable system to model the desired output. A well-known example of the latter is IBM's Watson, which handily beat the best human *Jeopardy!* players in 2011 but nevertheless gave a few answers that even novice human players would have known were wrong, such as the notorious "Toronto" answer to a Final Jeopardy question about "U.S. Cities."

In the context of TV entertainment, this was a harmless, amusing mistake—and a helpful reminder that even the smartest machines aren't perfect. Yet today Watson for Oncology is employed by more than a dozen cancer centers in the United States to "offer oncologists and people with cancer individualized treatment options" (IBM Watson 2016). Watson's diagnoses and treatment plans are still vetted by licensed oncologists. Still, how reliably can a human expert distinguish between a novel, unexpected treatment recommendation by Watson that might save a patient's life—something that has reportedly already happened in Japan (David 2016)—and the oncological equivalent of "Toronto"? At least in the context of oncology, a physician *can* take time to investigate and evaluate Watson's recommendations; but how can we insulate ourselves from the unpredictability of systems such as self-driving cars, in which the required speed of operation and decision-making may render real-time human supervision virtually impossible to implement?

Ideally, responsible creators of self-learning systems will allow them to operate "in the wild" only when their statistical failure rate in controlled settings is markedly lower than that of the average human performing the same task. Still, who should we hold responsible when a robot or other artificial agent *does* injure a person while honing its intelligence in the real world? Consider a catastrophic machine "error" that was not introduced by human programmers, could not have been specifically predicted by them, and thus could not have been prevented, except by not allowing the machine to act and learn in society in the first place.[7] What, if any, safeguards should be put in place to mitigate such losses, and who is responsible for making this happen? Lawmakers? Manufacturers? Individual AI scientists and programmers? Consumer groups? Insurance companies? We need a public conversation among affected stakeholders about what a *just distribution* of the risk burdens *and* benefits of self-learning systems will look like.

A related ethical issue concerns the very different degrees and types of risk that may be imposed by artificial agents on individuals and society. Allowing a self-driving security robot to patrol a mall food court risks bruising an errant toddler's foot, which is bad enough (Vincent 2016). It is quite another order of risk-magnitude to unleash a self-learning robot in a 2-ton metal chassis traveling public roads at highway speed, or to arm it with lethal weapons, or to link a self-learning agent up to critical power systems. Yet there are also significant risks involved with *not* employing self-learning systems, particularly in contexts such as driving and medicine where human error is a large and ineradicable source of grievous harms. If sound policy informed by careful ethical reflection does not begin to form soon around these questions of risk and responsibility in self-learning systems, the safety of innocent people *and* the long-term future of AI research may be gravely endangered.

22.3 Broader Ethical Concerns about Artificially Intelligent Robots

Not all ethical quandaries about AI-enabled robots are specific to their implementation of machine learning. Many such concerns apply to virtually any artificial agent capable of autonomous action in the world. These include such challenges as *meaningful human oversight and control of AI; algorithmic opacity and hidden machine bias; widespread technological unemployment; psychological and emotional manipulation of humans by AI*; and *automation bias.*

22.3.1 Meaningful Human Control and Oversight of AI

Society has an *ethical* interest in meaningful human control and oversight of AI, for several reasons. The first arises from the general ethical principle that humans are morally *responsible* for our chosen actions. Since, unlike our children, AI-enabled systems come into the world formed by deliberate human design, humans are in a deep sense always *morally accountable* for the effects of such agents on the world. It would therefore seem plainly irresponsible for humans to allow meaningful control or oversight of an artificial agent's actions to slip from our grasp.

A second reason for our ethical interest in meaningful human control and oversight of AI is its rapidly expanding scope of action. AI-enabled systems already operate in real-world contexts like driving and medicine that involve matters of life and death, as well as other core dimensions of human flourishing. Thus the effects of AI in the world for which humans are responsible—positive *and* negative—are of *increasing moral gravity*. This trend will strengthen as artificial systems demonstrate ever-greater competence and reliability in contexts with very high moral stakes (Wallach 2015).

As ethically fundamental as human responsibility for AI may be, the *practical* challenges of maintaining meaningful human control and oversight are immense. In addition to the aforementioned risk of emergent or other unpredictable AI behaviors, there are strong counter-pressures to *limit* human control and oversight of AI. Human supervisors are costly to employ, potentially reducing the profit to be reaped from automating a key task. Humans are also far slower to judge and act than are computers, so efficiency gains too can be diminished by our control. In many applications, such as driving, flight control, and financial trading, the entire function of the system will presuppose speeds and scales of decision-making beyond human reach. There is also the question of when our judgments warrant more *epistemic authority* or *epistemic trust* than machine judgments. If an artificially intelligent system has consistently demonstrated statistically greater competence than humans in a certain task, on what grounds do we give a human supervisor the power to challenge or override its decisions?

The difficulty is exacerbated by the fact that self-learning robots often operate in ways that are opaque to humans, even their programmers (Pasquale 2015). We must face the prospect of a growing disconnect between human and artificial forms of "expertise," a gap that should disturb us for several reasons. First, it risks the gradual devaluation of distinctly human skills and modes of understanding. Human expertise often expresses important moral and intellectual virtues missing from AI (such as perspective, empathy, integrity, aesthetic style, and civic-mindedness, to name a few)—virtues that are all too easily undervalued relative to AI's instrumental virtues of raw speed and efficiency. Additionally, productive AI–human *collaborations*—the chief goal of many researchers—will be far more difficult if AI and human agents cannot grasp one another's manner of reasoning, explain the basis of their decisions to one another, or pose critical questions to one another.

After all, if a human cannot reliably query an AI-enabled robot as to the specific evidence and chain of reasoning by means of which it arrived at its decision, how can he or she reliably assess the decision's validity? Human supervisors of AI agents cannot effectively do their job if their subordinate is a mute "black box." For this reason, many AI designers are looking for ways to increase the *transparency* of machine reasoning. For example, internal confidence measures reported alongside a given choice allow a human to give less credibility to decisions with a low confidence value. Still, the problem of *algorithmic opacity* remains a significant barrier to effective human oversight. It generates other moral risks as well.

22.3.2 Algorithmic Opacity and Hidden Machine Bias

In addition to frustrating meaningful human oversight of AI, "black boxed" or opaque algorithmic processes can perpetuate and reinforce morally and epistemically harmful biases. For example, racial, gender, or socioeconomic biases that originate in human minds are commonly embedded in the human-generated datasets used to train or "educate" machine systems. These data define the "world" that an artificial agent "knows." Yet the effect of human-biased data on machine outputs is easily obscured by several factors, making those biases more harmful and resistant to eradication.

One factor is algorithmic opacity itself. If I cannot know what features of a given dataset were singled out by a network's hidden layers as relevant and actionable, then I will be uncertain whether the network's decision rested on a harmful racial or gender bias encoded somewhere in that data. Another factor is our cultural tendency to think about robots and computers as inherently "objective" and "rational," and thus materially incapable of the kinds of emotional and psychological responses (e.g., fear, disgust, anger, shame) that typically produce irrational and harmful social biases. Even scientists who understand the mechanisms through which human bias can infect machine intelligence are often surprised to discover the extent of such bias in machine outputs—even from inputs thought to be relatively unbiased.

For example, a team of Boston University and Microsoft researchers found significant gender biases in machine "word embeddings" trained on a large body of Google News reports (Bolukbasi et al. 2016). They remark that data generated by "professional journalists" (as opposed to data sourced from internet message boards, for example) might have been expected to carry "little gender bias"; yet the machine outputs strongly reflected many harmful gender stereotypes (2016, 3). They observed that "the same system that solved [other] reasonable analogies will offensively answer 'man is to computer programmer as woman is to x' with x = *homemaker*" (3). The system also reflected "strong" racial stereotypes (15). To see how such biases could produce direct harm, just imagine this same system implemented in an AI agent tasked with providing college counseling to young men and women or with ranking employment applications for human resources managers at a large tech firm.

Indeed, hidden biases in AI algorithms and training data invite unjust outcomes or policies in predictive policing, lending, education, housing, healthcare, and employment, to name just a few sectors of AI implementation. Racial bias has already been found in facial recognition algorithms (Orcutt 2016) and, even more disturbingly, in machine-generated scores widely used by judges in criminal courts to predict the likelihood that a criminal defendant will reoffend (Angwin et al. 2016). Such scores shape judicial decisions about parole eligibility, length

of sentence, and the type of correctional facility to which a defendant will be subjected. A predictive engine that assigns higher risk scores to black defendants than to white defendants who are otherwise similar, and that systematically overestimates the likelihood of black recidivism while systematically *under*estimating the rate of recidivism among whites, not only reflects an existing social injustice, but *perpetuates* and *reinforces* it—both by giving it the stamp of machine objectivity and neutrality associated with computer-generated calculations and by encouraging further injustices (disproportionately longer and harsher sentences) against black defendants and their families.

Perhaps humans can learn to view biased machine algorithms and outputs with greater suspicion. Yet machine bias can also infect the judgments of less critically minded agents, such as robots tasked with identifying shoplifters, conducting anti-burglary patrols, or assisting with crowd-control or anti-terrorism operations. Such uses of robots are widely anticipated; indeed, automated security robots are already on the market (Vincent 2016). Perhaps we can train algorithms to *expose* hidden biases in such systems, for unless they can be effectively addressed, machine biases are virtually guaranteed to perpetuate and amplify many forms of social injustice.

22.3.3 Widespread Technological Unemployment

In the early nineteenth century, when weaving machines were introduced in England, great resentment and fear arose among textile workers who saw their jobs threatened. An organized revolt against the machines led by so-called Luddites (after a mythical hero known as "Ned Ludd" or "King Ludd") had sufficient cultural impact that, to this day, people who object to new developments in technology are known as "Neo-Luddites." Yet despite their very real harms, the disruptions of the Industrial Revolution produced social goods that not many would be willing to surrender for the old world: longer life spans, higher standards of living in industrialized nations, and, eventually, great expansions in skilled employment. The early computer revolution produced similar cultural disruptions, but a range of new public benefits and a booming market for jobs in the "knowledge economy."

Yet unlike other waves of machine automation, emerging advances in AI and robotics technology are now viewed as a significant threat to employees who perform *mental*, not just manual, labor. Automated systems already perform many tasks that traditionally required advanced education, such as legal discovery, reading x-ray films, grading essays, rating loan applications, and writing news articles (Kaplan 2015). IBM's Watson has been employed as a teaching assistant in an online college class on AI—without students discerning the non-human identity of their trusted TA "Jill" (Korn 2016).

The large-scale effects of AI and associated automation on human labor, social security, political stability, and economic equality are uncertain, but an Oxford study concludes that as many as 47% of U.S. jobs are at significant risk from advances in machine learning and mobile robotics (Frey and Osborne 2013, 1–2). Sectors at highest risk for displacement by automated systems include transportation and logistics (including driving jobs), sales, service, construction, office and administrative support, and production (Frey and Osborne 2013, 35). It is worth noting that the Oxford researchers were relatively conservative in their predictions of machine intelligence, suggesting that non-routine, high-skilled jobs associated with healthcare, scientific research, education, and the arts are at relatively low risk due to their heavy reliance on human creativity and social intelligence (Frey and Osborne 2013, 40). Yet more recent gains in machine learning have led many to anticipate a boom in artificial agents like the university TA "Jill Watson": able to compete with humans even in jobs that traditionally required social, creative, and intellectual capacities.

Such developments will profoundly challenge economic and political stability in a world already suffering from rising economic inequality, political disaffection, and growing class divisions. They also impact fundamental human values like autonomy and dignity, and make it even less certain that the benefits and risks of scientific and technical advances will be distributed among citizens and nations in a manner that is not merely efficient and productive, but also *good* and *just*.

22.3.4 *Psychological and Emotional Manipulation of Humans by AI*

The moral impacts of AI on human emotions, sociality, relationship bonding, public discourse, and civic character have only begun to be explored. Research in social AI for robots and other artificial agents is exploding, and vigorous efforts to develop carebots for the elderly, sexbots for the lonely, chatbots for customers and patients, and artificial assistants like Siri and Cortana for all of us are just the tip of the iceberg. The ethical questions that can arise in this domain are virtually limitless, since human sociality is the primary field of ethical action.

One deep worry about social AI is the well-documented tendency of humans to form robust emotional attachments to machines that simulate human emotional responses, even when the simulations are quite superficial (Turkle 2011). The behavior of artificial agents can also foster harmful delusions in humans, who may incorrectly perceive them as having human traits such as sentience, empathy, moral conscience, or loyalty. Thus humans are deeply vulnerable to emotional and psychological manipulation by AI and robotic systems coldly designed to exploit us for commercial, political, or other purposes (Scheutz 2012). Imagine,

for example, the public harm that could be done by a chatbot programmed to seek out and form emotionally manipulative online relationships with young voters or lonely seniors and then, once a bond is formed, to start interjecting deliberately manipulative messages about a political candidate whom the chatbot deeply "fears" or "loves."

Public education about the inability of robots to have feelings or form genuine bonds with humans will likely not be enough to prevent such harms, since even those with an insider's knowledge of the technology can find themselves responding to such powerful delusions (Scheutz 2012). It is thus imperative that lawmakers and developers of AI-enabled social agents begin to work together on ethical and legal guidelines for restricting or prohibiting harmful manipulation, particularly when it undermines human autonomy or damages our vital interests.

22.3.5 Automation Bias

A related ethical concern about human–robot/AI interaction is the psychological phenomenon of *automation bias*, in which humans greatly overestimate or rely unduly upon the capabilities of computerized systems (Cummings 2004). Automation bias can result from flawed expectations of computerized systems as infallible or inherently superior to human judgment, from time pressures or information overloads that make it difficult for humans to properly evaluate a computer's decision, or from overextending warranted confidence in a machine's actual capabilities into an area of action in which confidence is *not* warranted. The latter is often elicited on the basis of only shallow similarities with intelligent human behavior, as when pedestrians in Puerto Rico walked behind a self-parking Volvo in a garage, erroneously trusting the car (which lacked an optional "pedestrian-detection" package) to know not to back over a person (Hill 2015).

Automation bias has been cited as one possible factor in the 1988 downing of Iran Air 655 by the USS *Vincennes*, which caused the death of 290 civilians (Grut 2013; Galliott 2015, 217). Operators of the Aegis anti-aircraft system that mistakenly identified the airliner as a military jet had ample information to warrant overriding the identification, but failed to do so. As artificially intelligent and robotic systems are given increasing power to effect or incite action in the physical world, often with serious consequences for human safety and well-being, it is ethically imperative that the psychological dimension of human–AI and human–robot interactions be better understood. Such knowledge must guide efforts by AI and robotic system designers *and* users to reduce or eliminate harmful automation bias and other psychological misalignments between human interests and artificial cognition.

22.4. Public Fears and the Long-Term Future of AI

While many computer scientists consider AI to be simply an especially interesting aspect of their field, challenging to program, and sometimes frustrating if it does not behave as expected, in the popular press AI is frequently framed as a threat to humanity's survival. Elon Musk, founder and CEO of Space X and Tesla Motors, has warned the public that with AI we are "summoning the demon"; similar warnings about AI's *existential risks*, that is, its potential to threaten meaningful human existence, have been voiced by public figures such as Stephen Hawking and Bill Gates, along with a host of AI and robotics researchers (Standage 2016).[8] The urgency of their warnings is motivated by the unprecedented acceleration of developments in the field, especially in machine learning.[9]

Isaac Asimov's "Laws of Robotics" were an early fictional attempt to think through a set of guidelines for the control of intelligent robots (Asimov 2004). Today, however, technologists and ethicists must revisit this challenge in the face of profound public ambivalence about an AI-driven future. While many of us implicitly trust our car's directional sense or our tax software's financial acuity more than we trust our own, there is growing uncertainty about AI's long-term safety and compatibility with human interests.

Public distrust of AI systems could inhibit wider adoption and consumer engagement with these technologies, a consequence that AI researchers have good reason to want to avoid. Many public fears can be moderated by better education and communication from AI researchers about what AI today *is* (skillful at well-defined cognitive tasks) and is *not* (sentient, self-aware, malevolent *or* benevolent, or even robustly intelligent in the manner of humans). Moreover, we would all benefit from an outward expansion of public concern to the less apocalyptic, but more pressing ethical challenges of AI addressed in this chapter.

While talk about "Skynet" scenarios and "robot overlords" sounds like overheated speculation to many AI and robotics researchers—who are often overjoyed just to make a robot that can have a halfway convincing conversation or walk up an unstable hillside—growing public anxieties about AI and robotics technology may force researchers to pay more attention to such fears and at least begin an early dialogue about long-term control strategies for artificial agents. One does not have to predict a *Terminator*-like future to recognize that the ethical challenges presented by AI will not remain fixed in their present state; as the technology grows in power, complexity, and scale, so will its risks *and* benefits (Cameron 1984). For this reason, the ethics of artificial intelligence will be a rapidly moving target—and humanity as a whole must make a dedicated effort to keep up.

Notes

1. There is much debate (even between this chapter's authors) about whether computer science, and artificial intelligence research in particular, is a genuine science that studies and models natural phenomena (such as informational or computational processes) or is a branch of engineering.
2. An excellent introduction to the field appears in Russell and Norvig (2010).
3. IBM Watson, for example, prefers the term "augmented intelligence" or "cognitive computing" to "artificial intelligence," to emphasize Watson's potential to enhance and empower human intelligence, not render it superfluous (IBM Research 2016).
4. Research in AGI continues, if slowly, and is the central focus of organizations such as the Machine Intelligence Research Institute (MIRI). Many researchers who work on AGI are actively working to mitigate its considerable risks (Yudkowsky 2008).
5. A related study is that of *machine ethics*: designing agents with artificial *moral* intelligence. Because of space limitations, we restrict our focus here to the ethical challenges AI presents for *human* moral agents.
6. Depending on whether the emergent behavior is undesirable or useful to humans, emergence can be a "bug" (as with the 2010 Flash Crash caused by feedback loops among interacting global financial software systems) or a "feature" (as when an algorithm produces emergent and seemingly "intelligent" swarming behavior among a networked group of micro-robots).
7. Strictly speaking, machine learning networks do not make "errors"—they only generate unexpected or statistically rare outcomes that, from a human perspective, are not well aligned with programmers' or users' real-world goals for the system. But since human goals (for example, "promote human safety") are not actually *understood* by the system (however statistically effective it may be at reaching them), it cannot truly be said to "err" in producing a result incongruous with such a goal. The mistake, if there is one, is a gap or misalignment between what the machine's code and network weightings actually do and what its programmers wanted it to do. Good programming, training, and testing protocols can minimize such gaps, but it is virtually impossible to ensure that every such gap is eliminated.
8. See the open letter on AI from the Future of Life Institute (2005), with more than eight thousand signatories; see also Bostrom (2014).
9. Indeed, a standard textbook in AI, which twenty years ago had some three hundred pages, now, in its third edition, includes well over a thousand pages (Russell and Norvig 2010).

Works Cited

Achenbach, Joel. 2015. "Driverless Cars Are Colliding with the Creepy Trolley Problem." *Washington Post*, December 29. https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/.

Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. 2016. "Machine Bias." *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Asimov, Isaac. (1950) 2004. *I, Robot*. Reprint, New York: Bantam Books.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." *arXiv* 1607.06520v1 [cs.CL]. https://arxiv.org/abs/1607.06520.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Cameron, James (director). 1984. "The Terminator." Cameron, James and Gale Anne Hurd (writers).

Cummings, Mary L. 2004. "Automation Bias in Intelligent Time Critical Decision Support Systems." *AIAA 1st Intelligent Systems Technical Conference.* arc.aiaa.org/doi/pdf/10.2514/6.2004-6313.

David, Eric. 2016. "Watson Correctly Diagnoses Woman after Doctors Are Stumped." *SiliconAngle*, August 5. http://siliconangle.com/blog/2016/08/05/watson-correctly-diagnoses-woman-after-doctors-were-stumped/.

Future of Life Institute. 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence." http://futureoflife.org/ai-open-letter/.

Galliott, Jai. 2015. *Military Robots: Mapping the Moral Landscape*. New York: Routledge.

Grut, Chantal. 2013. "The Challenge of Autonomous Lethal Robotics to International Humanitarian Law." *Journal of Conflict and Security Law.* doi: 10.1093/jcsl/krt002. http://jcsl.oxfordjournals.org/content/18/1/5.abstract.

Hill, Kashmir. 2015. "Volvo Says Horrible 'Self-Parking Car Accident' Happened Because Driver Didn't Have 'Pedestrian Detection.'" *Fusion*, May 26. http://fusion.net/story/139703/self-parking-car-accident-no-pedestrian-detection/.

IBM Research. 2016. "Response to Request for Information: Preparing for the Future of Artificial Intelligence." https://www.research.ibm.com/cognitive-computing/ostp/rfi-response.shtml.

IBM Watson. 2016. "IBM Watson for Oncology." http://www.ibm.com/watson/watson-oncology.html.

Kaplan, Jerry. 2015. *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence*. New Haven, CT: Yale University Press.

Korn, Melissa. 2016. "Imagine Discovering That Your Teaching Assistant Really Is a Robot." *Wall Street Journal*, May 6. http://www.wsj.com/articles/if-your-teacher-sounds-like-a-robot-you-might-be-on-to-something-1462546621.

Moor, James. 2003. *The Turing Test: The Elusive Standard of Artificial Intelligence.* Dordrecht: Kluwer.

Orcutt, Mike. 2016. "Are Facial Recognition Systems Accurate? Depends on Your Race." *MIT Technology Review*, July 6. https://www.technologyreview.com/s/601786/are-face-recognition-systems-accurate-depends-on-your-race/.

Osborne, Michael and Carl Benedikt Frey. 2013. "The Future of Employment: How Susceptible Are Jobs to Computerisation?" *Oxford Martin Programme on the Impacts of Future Technology*. http://www.oxfordmartin.ox.ac.uk/publications/view/1314.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Russell, Stuart J. and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach.* 3d ed. Upper Saddle River, NJ: Pearson.

Scheutz, Matthias. 2012. "The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots." In *Robot Ethics*, edited by Patrick Lin, Keith Abney, and George Bekey, 205–222. Cambridge, MA: MIT Press.

Standage, Tom. 2016. "Artificial Intelligence: The Return of the Machinery Question." *Economist,* June 25. http://www.economist.com/news/special-report/21700761-after-many-false-starts-artificial-intelligence-has-taken-will-it-cause-mass.

Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind* 59: 236, 433–60.

Turkle, Sherry. 2011. *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Vincent, James. 2016. "Mall Security Bot Knocks Down Toddler, Breaks Asimov's First Law of Robotics." *Verge*, July 13. http://www.theverge.com/2016/7/13/12170640/mall-security-robot-k5-knocks-down-toddler.

Wallach, Wendell. 2015. *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control.* New York: Basic Books.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk," In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–45. New York: Oxford University Press.