CHAPTER 28

PHILOSOPHY, ETHICS, AND SAFETY OF AI

In which we consider the big questions around the meaning of AI, how we can ethically develop and apply it, and how we can keep it safe.

Philosophers have been asking big questions for a long time: How do minds work? Is it possible for machines to act intelligently in the way that people do? Would such machines have real, conscious minds?

To these, we add new ones: What are the ethical implications of intelligent machines in day-to-day use? Should machines be allowed to decide to kill humans? Can algorithms be fair and unbiased? What will humans do if machines can do all kinds of work? And how do we control machines that may become more intelligent than us?

28.1 The Limits of AI

Weak AI
Strong AI

In 1980, philosopher John Searle introduced a distinction between **weak AI**—the idea that machines could act *as if* they were intelligent—and **strong AI**—the assertion that machines that do so are *actually* consciously thinking (not just *simulating* thinking). Over time the definition of strong AI shifted to refer to what is also called "human-level AI" or "general AI"—programs that can solve an arbitrarily wide variety of tasks, including novel ones, and do so as well as a human.

Critics of weak AI who objected to the very possibility of intelligent behavior in machines now appear as shortsighted as Simon Newcomb, who in October 1903 wrote "aerial flight is one of the great class of problems with which man can never cope"—just two months before the Wright brothers' flight at Kitty Hawk. The rapid progress of recent years does not, however, prove that there can be no limits to what AI can achieve. Alan Turing (1950), the first person to define AI, was also the first to raise possible objections to AI, foreseeing almost all the ones subsequently raised by others.

28.1.1 The argument from informality

Turing's "argument from informality of behavior" says that human behavior is far too complex to be captured by any formal set of rules—humans must be using some informal guidelines that (the argument claims) could never be captured in a formal set of rules and thus could never be codified in a computer program.

A key proponent of this view was Hubert Dreyfus, who produced a series of influential critiques of artificial intelligence: *What Computers Can't Do* (1972), the sequel *What*

Computers Still Can't Do (1992), and, with his brother Stuart, *Mind Over Machine* (1986). Similarly, philosopher Kenneth Sayre (1993) said "Artificial intelligence *pursued within the cult of computationalism* stands not even a ghost of a chance of producing durable results." The technology they criticized came to be called **Good Old-Fashioned AI (GOFAI)**.

Good Old-Fashioned AI (GOFAI)

GOFAI corresponds to the simplest logical agent design described in Chapter 7, and we saw there that it is indeed difficult to capture every contingency of appropriate behavior in a set of necessary and sufficient logical rules; we called that the **qualification problem**. But as we saw in Chapter 12, probabilistic reasoning systems are more appropriate for open-ended domains, and as we saw in Chapter 22, deep learning systems do well on a variety of "informal" tasks. Thus, the critique is not addressed against computers *per se*, but rather against one particular style of programming them with logical rules—a style that was popular in the 1980s but has been eclipsed by new approaches.

One of Dreyfus's strongest arguments is for situated agents rather than disembodied logical inference engines. An agent whose understanding of "dog" comes only from a limited set of logical sentences such as "$Dog(x) \Rightarrow Mammal(x)$" is at a disadvantage compared to an agent that has watched dogs run, has played fetch with them, and has been licked by one. As philosopher Andy Clark (1998) says, "Biological brains are first and foremost the control systems for biological bodies. Biological bodies move and act in rich real-world surroundings." According to Clark, we are "good at frisbee, bad at logic."

The **embodied cognition** approach claims that it makes no sense to consider the brain separately: cognition takes place within a body, which is embedded in an environment. We need to study the system as a whole; the brain's functioning exploits regularities in its environment, including the rest of its body. Under the embodied cognition approach, robotics, vision, and other sensors become central, not peripheral.

Embodied cognition

Overall, Dreyfus saw areas where AI did not have complete answers and said that AI is therefore impossible; we now see many of these same areas undergoing continued research and development leading to increased capability, not impossibility.

28.1.2 The argument from disability

The "argument from disability" makes the claim that "a machine can never do *X*." As examples of *X*, Turing lists the following:

> Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as man, do something really new.

In retrospect, some of these are rather easy—we're all familiar with computers that "make mistakes." Computers with metareasoning capabilities (Chapter 6) can examine their own computations, thus being the subject of their own reasoning. A century-old technology has the proven ability to "make someone fall in love with it"—the teddy bear. Computer chess expert David Levy predicts that by 2050 people will routinely fall in love with humanoid robots. As for a robot falling in love, that is a common theme in fiction,[1] but there has been only limited academic speculation on the subject (Kim *et al.*, 2007). Computers have

[1] For example, the opera Coppélia (1870), the novel *Do Androids Dream of Electric Sheep?* (1968), the movies *AI* (2001), *Wall-E* (2008), and *Her* (2013).

done things that are "really new," making significant discoveries in astronomy, mathematics, chemistry, mineralogy, biology, computer science, and other fields, and creating new forms of art through style transfer (Gatys *et al.*, 2016). Overall, programs exceed human performance in some tasks and lag behind on others. The one thing that it is clear they can't do is be exactly human.

28.1.3 The mathematical objection

Turing (1936) and Gödel (1931) proved that certain mathematical questions are in principle unanswerable by particular formal systems. Gödel's incompleteness theorem (see Section 9.5) is the most famous example of this. Briefly, for any formal axiomatic framework F powerful enough to do arithmetic, it is possible to construct a so-called Gödel sentence $G(F)$ with the following properties:

- $G(F)$ is a sentence of F, but cannot be proved within F.
- If F is consistent, then $G(F)$ is true.

Philosophers such as J. R. Lucas (1961) have claimed that this theorem shows that machines are mentally inferior to humans, because machines are formal systems that are limited by the incompleteness theorem—they cannot establish the truth of their own Gödel sentence—while humans have no such limitation. This has caused a lot of controversy, spawning a vast literature, including two books by the mathematician/physicist Sir Roger Penrose (1989, 1994). Penrose repeats Lucas's claim with some fresh twists, such as the hypothesis that humans are different because their brains operate by quantum gravity—a theory that makes multiple false predictions about brain physiology.

We will examine three of the problems with Lucas's claim. First, an agent should not be ashamed that it cannot establish the truth of some sentence while other agents can. Consider the following sentence:

> Lucas cannot consistently assert that this sentence is true.

If Lucas asserted this sentence, then he would be contradicting himself, so therefore Lucas cannot consistently assert it, and hence it is true. We have thus demonstrated that there is a true sentence that Lucas cannot consistently assert while other people (and machines) can. But that does not make us think any less of Lucas.

Second, Gödel's incompleteness theorem and related results apply to *mathematics*, not to *computers*. No entity—human or machine—can prove things that are impossible to prove. Lucas and Penrose falsely assume that humans can somehow get around these limits, as when Lucas (1976) says "we must assume our own consistency, if thought is to be possible at all." But this is an unwarranted assumption: humans are notoriously inconsistent. This is certainly true for everyday reasoning, but it is also true for careful mathematical thought. A famous example is the four-color map problem. Alfred Kempe (1879) published a proof that was widely accepted for 11 years until Percy Heawood (1890) pointed out a flaw.

Third, Gödel's incompleteness theorem technically applies only to formal systems that are powerful enough to do arithmetic. This includes Turing machines, and Lucas's claim is in part based on the assertion that computers are equivalent to Turing machines. This is not quite true. Turing machines are infinite, whereas computers (and brains) are finite, and any computer can therefore be described as a (very large) system in propositional logic, which is not subject to Gödel's incompleteness theorem. Lucas assumes that humans can "change their

minds" while computers cannot, but that is also false—a computer can retract a conclusion after new evidence or further deliberation; it can upgrade its hardware; and it can change its decision-making processes with machine learning or software rewriting.

28.1.4 Measuring AI

Alan Turing, in his famous paper "Computing Machinery and Intelligence" (1950), suggested that instead of asking whether machines can think, we should ask whether machines can pass a behavioral test, which has come to be called the **Turing test**. The test requires a program to have a conversation (via typed messages) with an interrogator for five minutes. The interrogator then has to guess if the conversation is with a program or a person; the program passes the test if it fools the interrogator 30% of the time. To Turing, the key point was not the exact details of the test, but instead the idea of measuring intelligence by performance on some kind of open-ended behavioral task, rather than by philosophical speculation.

Nevertheless, Turing conjectured that by the year 2000 a computer with a storage of a billion units could pass the test, but here we are on the other side of 2000, and we still can't agree whether any program has passed. Many people have been fooled when they didn't know they might be chatting with a computer. The ELIZA program and Internet chatbots such as MGONZ (Humphrys, 2008) and NATACHATA (Jonathan *et al.*, 2009) fool their correspondents repeatedly, and the chatbot CYBERLOVER has attracted the attention of law enforcement because of its penchant for tricking fellow chatters into divulging enough personal information that their identity can be stolen.

In 2014, a chatbot called Eugene Goostman fooled 33% of the untrained amateur judges in a Turing test. The program claimed to be a boy from Ukraine with limited command of English; this helped explain its grammatical errors. Perhaps the Turing test is really a test of human gullibility. So far no well-trained judge has been fooled (Aaronson, 2014).

Turing test competitions have led to better chatbots, but have not been a focus of research within the AI community. Instead, AI researchers who crave competition are more likely to concentrate on playing chess or Go or StarCraft II, or taking an 8th grade science exam, or identifying objects in images. In many of these competitions, programs have reached or surpassed human-level performance, but that doesn't mean the programs are human-like outside the specific task. The point is to improve basic science and technology and to provide useful tools, not to fool judges.

28.2 Can Machines Really Think?

Some philosophers claim that a machine that acts intelligently would not be *actually* thinking, but would be only a *simulation* of thinking. But most AI researchers are not concerned with the distinction, and the computer scientist Edsger Dijkstra (1984) said that "The question of whether *Machines Can Think* . . . is about as relevant as the question of whether *Submarines Can Swim*." The American Heritage Dictionary's first definition of *swim* is "To move through water by means of the limbs, fins, or tail," and most people agree that submarines, being limbless, cannot swim. The dictionary also defines *fly* as "To move through the air by means of wings or winglike parts," and most people agree that airplanes, having winglike parts, can fly. However, neither the questions nor the answers have any relevance to the design or capabilities of airplanes and submarines; rather they are about word usage in English. (The

fact that ships do *swim* (*"privet"*) in Russian amplifies this point.) English speakers have not yet settled on a precise definition for the word "think"—does it require "a brain" or just "brain-like parts?"

Again, the issue was addressed by Turing. He notes that we never have *any* direct evidence about the internal mental states of other humans—a kind of mental solipsism. Nevertheless, Turing says, "Instead of arguing continually over this point, it is usual to have the **polite convention** that everyone thinks." Turing argues that we would also extend the polite convention to machines, if only we had experience with ones that act intelligently. However, now that we do have some experience, it seems that our willingness to ascribe sentience depends at least as much on humanoid appearance and voice as on pure intelligence.

Polite convention

28.2.1 The Chinese room

Chinese room

The philosopher John Searle rejects the polite convention. His famous **Chinese room** argument (Searle, 1990) goes as follows: Imagine a human, who understands only English, inside a room that contains a rule book, written in English, and various stacks of paper. Pieces of paper containing indecipherable symbols are slipped under the door to the room. The human follows the instructions in the rule book, finding symbols in the stacks, writing symbols on new pieces of paper, rearranging the stacks, and so on. Eventually, the instructions will cause one or more symbols to be transcribed onto a piece of paper that is passed back to the outside world. From the outside, we see a system that is taking input in the form of Chinese sentences and generating fluent, intelligent Chinese responses.

Searle then argues: it is given that the human does not understand Chinese. The rule book and the stacks of paper, being just pieces of paper, do not understand Chinese. Therefore, there is no understanding of Chinese. And Searle says that the Chinese room is doing the same thing that a computer would do, so therefore computers generate no understanding.

Biological naturalism

Searle (1980) is a proponent of **biological naturalism**, according to which mental states are high-level emergent features that are caused by low-level physical processes *in the neurons*, and it is the (unspecified) properties of the neurons that matter: according to Searle's biases, neurons have "it" and transistors do not. There have been many refutations of Searle's argument, but no consensus. His argument could equally well be used (perhaps by robots) to argue that a human cannot have true understanding; after all, a human is made out of cells, the cells do not understand, therefore there is no understanding. In fact, that is the plot of Terry Bisson's (1990) science fiction story *They're Made Out of Meat*, in which alien robots explore Earth and can't believe that hunks of meat could possibly be sentient. How they can be remains a mystery.

28.2.2 Consciousness and qualia

Consciousness

Running through all the debates about strong AI is the issue of **consciousness**: awareness of the outside world, and of the self, and the subjective experience of living. The technical term for the intrinsic nature of experiences is **qualia** (from the Latin word meaning, roughly, "of what kind"). The big question is whether machines can have qualia. In the movie *2001*, when astronaut David Bowman is disconnecting the "cognitive circuits" of the HAL 9000 computer, it says *"I'm afraid, Dave. Dave, my mind is going. I can feel it."* Does HAL actually have feelings (and deserve sympathy)? Or is the reply just an algorithmic response, no different from "Error 404: not found"?

Qualia

There is a similar question for animals: pet owners are certain that their dog or cat has consciousness, but not all scientists agree. Crickets change their behavior based on temperature, but few people would say that crickets experience the *feeling* of being warm or cold.

One reason that the problem of consciousness is hard is that it remains ill-defined, even after centuries of debate. But help may be on the way. Recently philosophers have teamed with neuroscientists under the auspices of the Templeton Foundation to start a series of experiments that could resolve some of the issues. Advocates of two leading theories of consciousness (global workspace theory and integrated information theory) have agreed that the experiments could confirm one theory over the other—a rarity in philosophy.

Alan Turing (1950) concedes that the question of consciousness is a difficult one, but denies that it has much relevance to the practice of AI: "I do not wish to give the impression that I think there is no mystery about consciousness ... But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper." We agree with Turing—we are interested in creating programs that behave intelligently. Individual aspects of consciousness—awareness, self-awareness, attention—can be programmed and can be part of an intelligent machine. The additional project of making a machine conscious in exactly the way humans are is not one that we are equipped to take on. We do agree that behaving intelligently will require some degree of *awareness*, which will differ from task to task, and that tasks involving interaction with humans will require a model of human subjective experience.

In the matter of modeling experience, humans have a clear advantage over machines, because they can use their own subjective apparatus to appreciate the subjective experience of others. For example, if you want to know *what it's like* when someone hits their thumb with a hammer, you can hit your thumb with a hammer. Machines have no such capability—although unlike humans, they can run each other's code.

28.3 The Ethics of AI

Given that AI is a powerful technology, we have a moral obligation to use it well, to promote the positive aspects and avoid or mitigate the negative ones.

The positive aspects are many. For example, AI can save lives through improved medical diagnosis, new medical discoveries, better prediction of extreme weather events, and safer driving with driver assistance and (eventually) self-driving technologies. There are also many opportunities to improve lives. Microsoft's AI for Humanitarian Action program applies AI to recovering from natural disasters, addressing the needs of children, protecting refugees, and promoting human rights. Google's AI for Social Good program supports work on rainforest protection, human rights jurisprudence, pollution monitoring, measurement of fossil fuel emissions, crisis counseling, news fact checking, suicide prevention, recycling, and other issues. The University of Chicago's Center for Data Science for Social Good applies machine learning to problems in criminal justice, economic development, education, public health, energy, and environment.

AI applications in crop management and food production help feed the world. Optimization of business processes using machine learning will make businesses more productive, increasing wealth and providing more employment. Automation can replace the tedious and dangerous tasks that many workers face, and free them to concentrate on more interesting

aspects. People with disabilities will benefit from AI-based assistance in seeing, hearing, and mobility. Machine translation already allows people from different cultures to communicate. Software-based AI solutions have near zero marginal cost of production, and so have the potential to democratize access to advanced technology (even as other aspects of software have the potential to centralize power).

Negative side effects

Despite these many positive aspects, we shouldn't ignore the negatives. Many new technologies have had unintended **negative side effects**: nuclear fission brought Chernobyl and the threat of global destruction; the internal combustion engine brought air pollution, global warming, and the paving of paradise. Other technologies can have negative effects even when used as intended, such as sarin gas, AR-15 rifles, and telephone solicitation. Automation will create wealth, but under current economic conditions much of that wealth will flow to the owners of the automated systems, leading to increased income inequality. This can be disruptive to a well-functioning society. In developing countries, the traditional path to growth through low-cost manufacturing for export may be cut off, as wealthy countries adopt fully automated manufacturing facilities on-shore. Our ethical and governance decisions will dictate the level of inequality that AI will engender.

All scientists and engineers face ethical considerations of what projects they should or should not take on, and how they can make sure the execution of the project is safe and beneficial. In 2010, the UK's Engineering and Physical Sciences Research Council held a meeting to develop a set of Principles of Robotics. In subsequent years other government agencies, nonprofit organizations, and companies created similar sets of principles. The gist is that every organization that creates AI technology, and everyone in the organization, has a responsibility to make sure the technology contributes to good, not harm. The most commonly-cited principles are:

Ensure safety	Establish accountability
Ensure fairness	Uphold human rights and values
Respect privacy	Reflect diversity/inclusion
Promote collaboration	Avoid concentration of power
Provide transparency	Acknowledge legal/policy implications
Limit harmful uses of AI	Contemplate implications for employment

Note that many of the principles, such as "ensure safety," have applicability to all software or hardware systems, not just AI systems. Several principles are worded in a vague way, making them difficult to measure or enforce. That is in part because AI is a big field with many subfields, each of which has a different set of historical norms and different relationships between the AI developers and the stakeholders. Mittelstadt (2019) suggests that the subfields should each develop more specific actionable guidelines and case precedents.

28.3.1 Lethal autonomous weapons

The UN defines a lethal autonomous weapon as one that locates, selects, and engages (i.e., kills) human targets without human supervision. Various weapons fulfill some of these criteria. For example, land mines have been used since the 17th century: they can select and engage targets in a limited sense according to the degree of pressure exerted or the quantity of metal present, but they cannot go out and locate targets by themselves. (Land mines are banned under the Ottawa Treaty.) Guided missiles, in use since the 1940s, can chase targets, but they have to be pointed in the right general direction by a human. Auto-firing

radar-controlled guns have been used to defend naval ships since the 1970s; they are mainly intended to destroy incoming missiles, but they could also attack manned aircraft. Although the word "autonomous" is often used to describe unmanned air vehicles or **drones**, most such weapons are both remotely piloted and require human actuation of the lethal payload.

At the time of writing, several weapons systems seem to have crossed the line into full autonomy. For example Israel's Harop missile is a "loitering munition" with a ten-foot wingspan and a fifty-pound warhead. It searches for up to six hours in a given geographical region for any target that meets a given criterion and then destroys it. The criterion could be "emits a radar signal resembling antiaircraft radar" or "looks like a tank." The Turkish manufacturer STM advertises its Kargu quadcopter—which carries up to 1.5kg of explosives—as capable of "Autonomous hit . . . targets selected on images . . . tracking moving targets . . . anti-personnel . . . face recognition."

Autonomous weapons have been called the "third revolution in warfare" after gunpowder and nuclear weapons. Their military potential is obvious. For example, few experts doubt that autonomous fighter aircraft would defeat any human pilot. Autonomous aircraft, tanks, and submarines can be cheaper, faster, more maneuverable, and have longer range than their manned counterparts.

Since 2014, the United Nations in Geneva has conducted regular discussions under the auspices of the Convention on Certain Conventional Weapons (CCW) on the question of whether to ban lethal autonomous weapons. At the time of writing, 30 nations, ranging in size from China to the Holy See, have declared their support for an international treaty, while other key countries—including Israel, Russia, South Korea, and the United States—are opposed to a ban.

The debate over autonomous weapons includes legal, ethical and practical aspects. The legal issues are governed primarily by the CCW, which requires the possibility of discriminating between combatants and non-combatants, the judgment of military necessity for an attack, and the assessment of proportionality between the military value of a target and the possibility of collateral damage. The feasibility of meeting these criteria is an engineering question—one whose answer will undoubtedly change over time. At present, discrimination seems feasible in some circumstances and will undoubtedly improve rapidly, but necessity and proportionality are not presently feasible: they require that machines make subjective and situational judgments that are considerably more difficult than the relatively simple tasks of searching for and engaging potential targets. For these reasons, it would be legal to use autonomous weapons only in circumstances where a human operator can reasonably predict that the execution of the mission will not result in civilians being targeted or the weapons conducting unnecessary or disproportionate attacks. This means that, for the time being, only very restricted missions could be undertaken by autonomous weapons.

On the ethical side, some find it simply morally unacceptable to delegate the decision to kill humans to a machine. For example, Germany's ambassador in Geneva has stated that it "will not accept that the decision over life and death is taken solely by an autonomous system" while Japan "has no plan to develop robots with humans out of the loop, which may be capable of committing murder." Gen. Paul Selva, at the time the second-ranking military officer in the United States, said in 2017, "I don't think it's reasonable for us to put robots in charge of whether or not we take a human life." Finally, António Guterres, the head of the United Nations, stated in 2019 that "machines with the power and discretion to take lives

without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law."

More than 140 NGOs in over 60 countries are part of the Campaign to Stop Killer Robots, and an open letter organized in 2015 by the Future of Life Institute organized an open letter was signed by over 4,000 AI researchers[2] and 22,000 others.

Against this, it can be argued that as technology improves it ought to be possible to develop weapons that are *less* likely than human soldiers or pilots to cause civilian casualties. (There is also the important benefit that autonomous weapons reduce the need for human soldiers and pilots to risk death.) Autonomous systems will not succumb to fatigue, frustration, hysteria, fear, anger, or revenge, and need not "shoot first, ask questions later" (Arkin, 2015). Just as guided munitions have reduced collateral damage compared to unguided bombs, one may expect intelligent weapons to further improve the precision of attacks. (Against this, see Benjamin (2013) for an analysis of drone warfare casualties.) This, apparently, is the position of the United States in the latest round of negotiations in Geneva.

Perhaps counterintuitively, the United States is also one of the few nations whose own policies currently preclude the use of autonomous weapons. The 2011 U.S. Department of Defense (DOD) roadmap says: "For the foreseeable future, decisions over the use of force [by autonomous systems] and the choice of which individual targets to engage with lethal force will be retained under human control." The primary reason for this policy is practical: autonomous systems are not reliable enough to be trusted with military decisions.

The issue of reliability came to the fore on September 26, 1983, when Soviet missile officer Stanislav Petrov's computer display flashed an alert of an incoming missile attack. According to protocol, Petrov should have initiated a nuclear counterattack, but he suspected the alert was a bug and treated it as such. He was correct, and World War III was (narrowly) averted. We don't know what would have happened if there had been no human in the loop.

Reliability is a very serious concern for military commanders, who know well the complexity of battlefield situations. Machine learning systems that operate flawlessly in training may perform poorly when deployed. Cyberattacks against autonomous weapons could result in friendly-fire casualties; disconnecting the weapon from all communication may prevent that (assuming it has not already been compromised), but then the weapon cannot be recalled if it is malfunctioning.

The overriding practical issue with autonomous weapons is that they they are scalable weapons of mass destruction, in the sense that the scale of an attack that can be launched is proportional to the amount of hardware one can afford to deploy. A quadcopter two inches in diameter can carry a lethal explosive charge, and one million can fit in a regular shipping container. Precisely because they are autonomous, these weapons would not need one million human supervisors to do their work.

As weapons of mass destruction, scalable autonomous weapons have advantages for the attacker compared to nuclear weapons and carpet bombing: they leave property intact and can be applied selectively to eliminate only those who might threaten an occupying force. They could certainly be used to wipe out an entire ethnic group or all the adherents of a particular religion. In many situations, they would also be untraceable. These characteristics make them particularly attractive to non-state actors.

[2] Including the two authors of this book.

These considerations—particularly those characteristics that advantage the attacker—suggest that autonomous weapons will reduce global and national security for all parties. The rational response for governments seems to be to engage in arms control discussions rather than an arms race.

The process of designing a treaty is not without its difficulties, however. AI is a **dual use** technology: AI technologies that have peaceful applications such as flight control, visual tracking, mapping, navigation, and multiagent planning, can easily be applied to military purposes. It is easy to turn an autonomous quadcopter into a weapon simply by attaching an explosive and commanding it to seek out a target. Dealing with this will require careful implementation of compliance regimes with industry cooperation, as has already been demonstrated with some success by the Chemical Weapons Convention.

Dual use

28.3.2 Surveillance, security, and privacy

In 1976, Joseph Weizenbaum warned that automated speech recognition technology could lead to widespread wiretapping, and hence to a loss of civil liberties. Today, that threat has been realized, with most electronic communication going through central servers that can be monitored, and cities packed with microphones and cameras that can identify and track individuals based on their voice, face, and gait. Surveillance that used to require expensive and scarce human resources can now be done at a mass scale by machines.

As of 2018, there were as many as 350 million **surveillance cameras** in China and 70 million in the United States. China and other countries have begun exporting surveillance technology to low-tech countries, some with reputations for mistreating their citizens and disproportionately targeting marginalized communities. AI engineers should be clear on what uses of surveillance are compatible with human rights, and decline to work on applications that are incompatible.

Surveillance camera

As more of our institutions operate online, we become more vulnerable to cybercrime (phishing, credit card fraud, botnets, ransomware) and cyberterrorism (including potentially deadly attacks such as shutting down hospitals and power plants or commandeering self-driving cars). Machine learning can be a powerful tool for both sides in the **cybersecurity** battle. Attackers can use automation to probe for insecurities and they can apply reinforcement learning for phishing attempts and automated blackmail. Defenders can use unsupervised learning to detect anomalous incoming traffic patterns (Chandola *et al.*, 2009; Malhotra *et al.*, 2015) and various machine learning techniques to detect fraud (Fawcett and Provost, 1997; Bolton and Hand, 2002). As attacks get more sophisticated, there is a greater responsibility for all engineers, not just the security experts, to design secure systems from the start. One forecast (Kanal, 2017) puts the market for machine learning in cybersecurity at about $100 billion by 2021.

Cybersecurity

As we interact with computers for increasing amounts of our daily lives, more data on us is being collected by governments and corporations. Data collectors have a moral and legal responsibility to be good stewards of the data they hold. In the U.S., the Health Insurance Portability and Accountability Act (HIPAA) and the Family Educational Rights and Privacy Act (FERPA) protect the privacy of medical and student records. The European Union's General Data Protection Regulation (GDPR) mandates that companies design their systems with protection of data in mind and requires that they obtain user consent for any collection or processing of data.

Balanced against the individual's right to privacy is the value that society gains from sharing data. We want to be able to stop terrorists without oppressing peaceful dissent, and we want to cure diseases without compromising any individual's right to keep their health history private. One key practice is **de-identification**: eliminating personally identifying information (such as name and social security number) so that medical researchers can use the data to advance the common good. The problem is that the shared de-identified data may be subject to re-identification. For example, if the data strips out the name, social security number, and street address, but includes date of birth, gender, and zip code, then, as shown by Latanya Sweeney (2000), 87% of the U.S. population can be uniquely re-identified. Sweeney emphasized this point by re-identifying the health record for the governor of her state when he was admitted to the hospital. In the **Netflix Prize** competition, de-identified records of individual movie ratings were released, and competitors were asked to come up with a machine learning algorithm that could accurately predict which movies an individual would like. But researchers were able to re-identify individual users by matching the date of a rating in the Netflix database with the date of a similar ranking in the Internet Movie Database (IMDB), where users sometimes use their actual names (Narayanan and Shmatikov, 2006).

De-identification

Netflix Prize

This risk can be mitigated somewhat by **generalizing fields**: for example, replacing the exact birth date with just the year of birth, or a broader range like "20-30 years old." Deleting a field altogether can be seen as a form of generalizing to "any." But generalization alone does not guarantee that records are safe from re-identification; it may be that there is only one person in zip code 94720 who is 90–100 years old. A useful property is **k-anonymity**: a database is k-anonymized if every record in the database is indistinguishable from at least $k-1$ other records. If there are records that are more unique than this, they would have to be further generalized.

K-anonymity

An alternative to sharing de-identified records is to keep all records private, but allow **aggregate querying**. An API for queries against the database is provided, and valid queries receive a response that summarizes the data with a count or average. But no response is given if it would violate certain guarantees of privacy. For example, we could allow an epidemiologist to ask, for each zip code, the percentage of people with cancer. For zip codes with at least n people a percentage would be given (with a small amount of random noise), but no response would be given for zip codes with fewer than n people..

Aggregate querying

Care must be taken to protect against de-identification using multiple queries. For example, if the query "average salary and number of employees of XYZ company age 30-40" gives the response [$81,234, 12] and the query "average salary and number of employees of XYZ company age 30-41" gives the response [$81,199, 13], and if we use LinkedIn to find the one 41-year-old at XYZ company, then we have successfully identified them, and can compute their exact salary, even though all the responses involved 12 or more people. The system must be carefully designed to protect against this, with a combination of limits on the queries that can be asked (perhaps only a predefined set of non-overlapping age ranges can be queried) and the precision of the results (perhaps both queries give the answer "about $81,000").

A stronger guarantee is **differential privacy**, which assures that an attacker cannot use queries to re-identify any individual in the database, even if the attacker can make multiple queries and has access to separate linking databases. The query response employs a randomized algorithm that adds a small amount of noise to the result. Given a database D, any record in the database r, any query Q, and a possible response y to the query, we say that the database

Differential privacy

D has ϵ–differential privacy if the log probability of the response y varies by less than ϵ when we add the record r:

$$|\log P(Q(D)=y) - \log P(Q(D+r)=y)| \leq \epsilon\,.$$

In other words, whether any one person decides to participate in the data base or not makes no appreciable difference to the answers anyone can get, and therefore there is no privacy disincentive to participate. Many databases are designed to guarantee differential privacy.

So far we have considered the issue of sharing de-identified data from a central database. An approach called **federated learning** (Konečnỳ *et al.*, 2016) has no central database; instead, users maintain their own local databases that keep their data private. However, they can share parameters of a machine learning model that is enhanced with their data, without the risk of revealing any of the private data. Imagine a speech understanding application that users can run locally on their phone. The application contains a baseline neural network, which is then improved by local training on the words that are heard on the user's phone. Periodically, the owners of the application poll a subset of the users and ask them for the parameter values of their improved local network, but not for any of their raw data. The parameter values are combined together to form a new improved model which is then made available to all users, so that they all get the benefit of the training that is done by other users.

Federated learning

For this scheme to preserve privacy, we have to be able to guarantee that the model parameters shared by each user cannot be reverse-engineered. If we sent the raw parameters, there is a chance that an adversary inspecting them could deduce whether, say, a certain word had been heard by the user's phone. One way to eliminate this risk is with **secure aggregation** (Bonawitz *et al.*, 2017). The idea is that the central server doesn't need to know the exact parameter value from each distributed user; it only needs to know the average value for each parameter, over all polled users. So each user can disguise their parameter values by adding a unique mask to each value; as long as the sum of the masks is zero, the central server will be able to compute the correct average. Details of the protocol make sure that it is efficient in terms of communication (less than half the bits transmitted correspond to masking), is robust to individual users failing to respond, and is secure in the face of adversarial users, eavesdroppers, or even an adversarial central server.

Secure aggregation

28.3.3 Fairness and bias

Machine learning is augmenting and sometimes replacing human decision-making in important situations: whose loan gets approved, to what neighborhoods police officers are deployed, who gets pretrial release or parole. But machine learning models can perpetuate **societal bias**. Consider the example of an algorithm to predict whether criminal defendants are likely to re-offend, and thus whether they should be released before trial. It could well be that such a system picks up the racial or gender prejudices of human judges from the examples in the training set. Designers of machine learning systems have a moral responsibility to ensure that their systems are in fact fair. In regulated domains such as credit, education, employment, and housing, they have a legal responsibility as well. But what is fairness? There are multiple criteria; here are six of the most commonly-used concepts:

Societal bias

- **Individual fairness**: A requirement that individuals are treated similarly to other similar individuals, regardless of what class they are in.

- **Group fairness**: A requirement that two classes are treated similarly, as measured by some summary statistic.
- **Fairness through unawareness**: If we delete the race and gender attributes from the data set, then it might seem that the system cannot discriminate on those attributes. Unfortunately, we know that machine learning models can predict latent variables (such as race and gender) given other correlated variables (such as zip code and occupation). Furthermore, deleting those attributes makes it impossible to verify equal opportunity or equal outcomes. Still, some countries (e.g., Germany) have chosen this approach for their demographic statistics (whether or not machine learning models are involved).
- **Equal outcome**: The idea that each demographic class gets the same results; they have **demographic parity**. For example, suppose we have to decide whether we should approve loan applications; the goal is to approve those applicants who will pay back the loan and not those who will default on the loan. Demographic parity says that both males and females should have the same percentage of loans approved. Note that this is a group fairness criterion that does nothing to ensure individual fairness; a well-qualified applicant might be denied and a poorly-qualified applicant might be approved, as long as the overall percentages are equal. Also, this approach favors redress of past biases over accuracy of prediction. If a man and a woman are equal in every way, except the woman receives a lower salary for the same job, should she be approved because she would be equal if not for historical biases, or should she be denied because the lower salary does in fact make her more likely to default?

Demographic parity

- **Equal opportunity**: The idea that the people who truly have the ability to pay back the loan should have an equal chance of being correctly classified as such, regardless of their sex. This approach is also called "balance." It can lead to unequal outcomes and ignores the effect of bias in the societal processes that produced the training data.
- **Equal impact**: People with similar likelihood to pay back the loan should have the same expected utility, regardless of the class they belong to. This goes beyond equal opportunity in that it considers both the benefits of a true prediction and the costs of a false prediction.

Let us examine how these issues play out in a particular context. COMPAS is a commercial system for recidivism (re-offense) scoring. It assigns to a defendant in a criminal case a **risk score**, which is then used by a judge to help make decisions: Is it safe to release the defendant before trial, or should they be held in jail? If convicted, how long should the sentence be? Should parole be granted? Given the significance of these decisions, the system has been the subject of intense scrutiny (Dressel and Farid, 2018).

Well calibrated

COMPAS is designed to be **well calibrated**: all the individuals who are given the same score by the algorithm should have approximately the same probability of re-offending, regardless of race. For example, among all people that the model assigns a risk score of 7 out of 10, 60% of whites and 61% of blacks re-offend. The designers thus claim that it meets the desired fairness goal.

On the other hand, COMPAS does not achieve equal opportunity: the proportion of those who did not re-offend but were falsely rated as high-risk was 45% for blacks and 23% for whites. In the case *State v. Loomis*, where a judge relied on COMPAS to determine the sentence of the defendant, Loomis argued that the secretive inner workings of the algorithm

violated his due process rights. Though the Wisconsin Supreme Court found that the sentence given would be no different without COMPAS in this case, it did issue warnings about the algorithm's accuracy and risks to minority defendants. Other researchers have questioned whether it is appropriate to use algorithms in applications such as sentencing.

We could hope for an algorithm that is both well calibrated and equal opportunity, but, as Kleinberg *et al.* (2016) show, that is impossible. If the base classes are different, then any algorithm that is well calibrated will necessarily not provide equal opportunity, and vice versa. How can we weigh the two criteria? Equal impact is one possibility. In the case of COMPAS, this means weighing the negative utility of defendants being falsely classified as high risk and losing their freedom, versus the cost to society of an additional crime being committed, and finding the point that optimizes the tradeoff. This is complicated because there are multiple costs to consider. There are individual costs—a defendant who is wrongfully held in jail suffers a loss, as does the victim of a defendant who was wrongfully released and re-offends. But beyond that there are group costs—everyone has a certain fear that they will be wrongfully jailed, or will be the victim of a crime, and all taxpayers contribute to the costs of jails and courts. If we give value to those fears and costs in proportion to the size of a group, then utility for the majority may come at the expense of a minority.

Another problem with the whole idea of recidivism scoring, regardless of the model used, is that we don't have unbiased ground truth data. The data does not tell us who has *committed* a crime—all we know is who has been *convicted* of a crime. If the arresting officers, judge, or jury is biased, then the data will be biased. If more officers patrol some locations, then the data will be biased against people in those locations. Only defendants who are released are candidates to recommit, so if the judges making the release decisions are biased, the data may be biased. If you assume that behind the biased data set there is an underlying, unknown, unbiased data set which has been corrupted by an agent with biases, then there are techniques to recover an approximation to the unbiased data. Jiang and Nachum (2019) describe various scenarios and the techniques involved.

One more risk is that machine learning can be used to *justify* bias. If decisions are made by a biased human after consulting with a machine learning system, the human can say "here is how my interpretation of the model supports my decision, so you shouldn't question my decision." But other interpretations could lead to an opposite decision.

Sometimes fairness means that we should reconsider the objective function, not the data or the algorithm. For example, in making job hiring decisions, if the objective is to hire candidates with the best qualifications in hand, we risk unfairly rewarding those who have had advantageous educational opportunities throughout their lives, thereby enforcing class boundaries. But if the objective is to hire candidates with the best ability to learn on the job, we have a better chance to cut across class boundaries and choose from a broader pool. Many companies have programs designed for such applicants, and find that after a year of training, the employees hired this way do as well as the traditional candidates. Similarly, just 18% of computer science graduates in the U.S. are women, but some schools, such as Harvey Mudd University, have achieved 50% parity with an approach that is focused on encouraging and retaining those who start the computer science program, especially those who start with less programming experience.

A final complication is deciding which classes deserve protection. In the U.S., the Fair Housing Act recognized seven protected classes: race, color, religion, national origin, sex,

disability, and familial status. Other local, state, and federal laws recognize other classes, including sexual orientation, and pregnancy, marital, and veteran status. Is it fair that these classes count for some laws and not others? International human rights law, which encompasses a broad set of protected classes, is a potential framework to harmonize protections across various groups.

Sample size disparity

Even in the absence of societal bias, **sample size disparity** can lead to biased results. In most data sets there will be fewer training examples of minority class individuals than of majority class individuals. Machine learning algorithms give better accuracy with more training data, so that means that members of minority classes will experience lower accuracy. For example, Buolamwini and Gebru (2018) examined a computer vision gender identification service, and found that it had near-perfect accuracy for light-skinned males, and a 33% error rate for dark-skinned females. A constrained model may not be able to simultaneously fit both the majority and minority class—a linear regression model might minimize average error by fitting just the majority class, and in an SVM model, the support vectors might all correspond to majority class members.

Bias can also come into play in the software development process (whether or not the software involves machine learning). Engineers who are debugging a system are more likely to notice and fix those problems that are applicable to themselves. For example, it is difficult to notice that a user interface design won't work for colorblind people unless you are in fact colorblind, or that an Urdu language translation is faulty if you don't speak Urdu.

How can we defend against these biases? First, understand the limits of the data you are using. It has been suggested that data sets (Gebru *et al.*, 2018; Hind *et al.*, 2018) and models (Mitchell *et al.*, 2019) should come with annotations: declarations of provenance, security, conformity, and fitness for use. This is similar to the **data sheets** that accompany electronic components such as resistors; they allow designers to decide what components to use. In addition to the data sheets, it is important to train engineers to be aware of issues of fairness and bias, both in school and with on-the-job training. Having a diversity of engineers from different backgrounds makes it easier for them to notice problems in the data or models. A study by the AI Now Institute (West *et al.*, 2019) found that only 18% of authors at leading AI conferences and 20% of AI professors are women. Black AI workers are at less than 4%. Rates at industry research labs are similar. Diversity could be increased by programs earlier in the pipeline—in college or high school—and by greater awareness at the professional level. Joy Buolamwini founded the Algorithmic Justice League to raise awareness of this issue and develop practices for accountability.

Data sheet

A second idea is to de-bias the data (Zemel *et al.*, 2013). We could over-sample from minority classes to defend against sample size disparity. Techniques such as SMOTE, the synthetic minority over-sampling technique (Chawla *et al.*, 2002) or ADASYN, the adaptive synthetic sampling approach for imbalanced learning (He *et al.*, 2008), provide principled ways of oversampling. We could examine the provenance of data and, for example, eliminate examples from judges who have exhibited bias in their past court cases. Some analysts object to the idea of discarding data, and instead would recommend building a hierarchical model of the data that includes sources of bias, so they can be modeled and compensated for. Google and NeurIPS have attempted to raise awareness of this issue by sponsoring the Inclusive Images Competition, in which competitors train a network on a data set of labeled images collected in North America and Europe, and then test it on images taken from all around the

world. The issue is that given this data set, it is easy to apply the label "bride" to a woman in a standard Western wedding dress, but harder to recognize traditional African and Indian matrimonial dress.

A third idea is to invent new machine learning models and algorithms that are more resistant to bias; and the final idea is to let a system make initial recommendations that may be biased, but then train a second system to de-bias the recommendations of the first one. Bellamy *et al.* (2018) introduced the IBM AI FAIRNESS 360 system, which provides a framework for all of these ideas. We expect there will be increased use of tools like this in the future.

How do you make sure that the systems you build will be fair? A set of best practices has been emerging (although they are not always followed):

- Make sure that the software engineers talk with social scientists and domain experts to understand the issues and perspectives, and consider fairness from the start.
- Create an environment that fosters the development of a diverse pool of software engineers that are representative of society.
- Define what groups your system will support: different language speakers, different age groups, different abilities with sight and hearing, etc.
- Optimize for an objective function that incorporates fairness.
- Examine your data for prejudice and for correlations between protected attributes and other attributes.
- Understand how any human annotation of data is done, design goals for annotation accuracy, and verify that the goals are met.
- Don't just track overall metrics for your system; make sure you track metrics for subgroups that might be victims of bias.
- Include system tests that reflect the experience of minority group users.
- Have a feedback loop so that when fairness problems come up, they are dealt with.

28.3.4 Trust and transparency

Trust

It is one challenge to make an AI system accurate, fair, safe, and secure; a different challenge to convince everyone else that you have done so. People need to be able to **trust** the systems they use. A PwC survey in 2017 found that 76% of businesses were slowing the adoption of AI because of trustworthiness concerns. In Section 19.9.4 we covered some of the engineering approaches to trust; here we discuss the policy issues.

Verification and validation

To earn trust, any engineered systems must go through a **verification and validation** (V&V) process. Verification means that the product satisfies the specifications. Validation means ensuring that the specifications actually meet the needs of the user and other affected parties. We have an elaborate V&V methodology for engineering in general, and for traditional software development done by human coders; much of that is applicable to AI systems. But machine learning systems are different and demand a different V&V process, which has not yet been fully developed. We need to verify the data that these systems learn from; we need to verify the accuracy and fairness of the results, even in the face of uncertainty that makes an exact result unknowable; and we need to verify that adversaries cannot unduly influence the model, nor steal information by querying the resulting model.

Certification

One instrument of trust is **certification**; for example, Underwriters Laboratories (UL) was founded in 1894 at a time when consumers were apprehensive about the risks of electric

power. UL certification of appliances gave consumers increased trust, and in fact UL is now considering entering the business of product testing and certification for AI.

Other industries have long had safety standards. For example, ISO 26262 is an international standard for the safety of automobiles, describing how to develop, produce, operate, and service vehicles in a safe way. The AI industry is not yet at this level of clarity, although there are some frameworks in progress, such as IEEE P7001, a standard defining ethical design for artificial intelligence and autonomous systems (Bryson and Winfield, 2017). There is ongoing debate about what kind of certification is necessary, and to what extent it should be done by the government, by professional organizations like IEEE, by independent certifiers such as UL, or through self-regulation by the product companies.

Transparency

Another aspect of trust is **transparency**: consumers want to know what is going on inside a system, and that the system is not working against them, whether due to intentional malice, an unintentional bug, or pervasive societal bias that is recapitulated by the system. In some cases this transparency is delivered directly to the consumer. In other cases their are intellectual property issues that keep some aspects of the system hidden to consumers, but open to regulators and certification agencies.

When an AI system turns you down for a loan, you deserve an explanation. In Europe, the GDPR enforces this for you. An AI system that can explain itself is called **explainable AI (XAI)**. A good explanation has several properties: it should be understandable and convincing to the user, it should accurately reflect the reasoning of the system, it should be complete, and it should be specific in that different users with different conditions or different outcomes should get different explanations.

Explainable AI (XAI)

It is quite easy to give a decision algorithm access to its own deliberative processes, simply by recording them and making them available as data structures. This means that machines may eventually be able to give better explanations of their decisions than humans can. Moreover, we can take steps to certify that the machine's explanations are not deceptions (intentional or self-deception), something that is more difficult with a human.

An explanation is a helpful but not sufficient ingredient to trust. One issue is that explanations are not decisions: they are stories about decisions. As discussed in Section 19.9.4, we say that a system is interpretable if we can inspect the source code of the model and see what it is doing, and we say it is explainable if we can make up a story about what it is doing—even if the system itself is an uninterpretable black box. To explain an uninterpretable black box, we need to build, debug, and test a separate explanation system, and make sure it is in sync with the original system. And because humans love a good story, we are all too willing to be swayed by an explanation that sounds good. Take any political controversy of the day, and you can always find two so-called experts with diametrically opposed explanations, both of which are internally consistent.

A final issue is that an explanation about one case does not give you a summary over other cases. If the bank explains, "Sorry, you didn't get the loan because you have a history of previous financial problems," you don't know if that explanation is accurate or if the bank is secretly biased against you for some reason. In this case, you require not just an explanation, but also an **audit** of past decisions, with aggregated statistics across various demographic groups, to see if their approval rates are balanced.

Part of transparency is knowing whether you are interacting with an AI system or a human. Toby Walsh (2015) proposed that "an autonomous system should be designed so that

it is unlikely to be mistaken for anything besides an autonomous system, and should identify itself at the start of any interaction." He called this the "red flag" law, in honor of the UK's 1865 Locomotive Act, which required any motorized vehicle to have a person with a red flag walk in front of it, to signal the oncoming danger.

In 2019, California enacted a law stating that "It shall be unlawful for any person to use a bot to communicate or interact with another person in California online, with the intent to mislead the other person about its artificial identity."

28.3.5 The future of work

From the first agricultural revolution (10,000 BCE) to the industrial revolution (late 18th century) to the green revolution in food production (1950s), new technologies have changed the way humanity works and lives. A primary concern arising from the advance of AI is that human labor will become obsolete. Aristotle, in Book I of his *Politics*, presents the main point quite clearly:

> For if every instrument could accomplish its own work, obeying or anticipating the will of others . . . if, in like manner, the shuttle would weave and the plectrum touch the lyre without a hand to guide them, chief workmen would not want servants, nor masters slaves.

Everyone agrees with Aristotle's observation that there is an immediate reduction in employment when an employer finds a mechanical method to perform work previously done by a person. The issue is whether the so-called compensation effects that ensue—and that tend to increase employment—will eventually make up for this reduction. The primary compensation effect is the increase in overall wealth from greater productivity, which leads in turn to greater demand for goods and tends to increase employment. For example, PwC (Rao and Verweij, 2017) predicts that AI contribute $15 trillion annually to global GDP by 2030. The healthcare and automotive/transportation industries stand to gain the most in the short term. However, the advantages of automation have not yet taken over in our economy: the current rate of growth in labor productivity is actually below historical standards. Brynjolfsson *et al.* (2018) attempt to explain this paradox by suggesting that the lag between the development of basic technology and its implementation in the economy is longer than commonly supposed.

Technological innovations have historically put some people out of work. Weavers were replaced by automated looms in the 1810s, leading to the Luddite protests. The Luddites were not against technology *per se*; they just wanted the machines to be used by skilled workers paid a good wage to make high-quality goods, rather than by unskilled workers to make poor-quality goods at low wages. The global destruction of jobs in the 1930s led John Maynard Keynes to coin the term **technological unemployment**. In both cases, and several others, employment levels eventally recovered.

Technological unemployment

The mainstream economic view for most of the 20th century was that technological employment was at most a short-term phenomenon. Increased productivity would always lead to increased wealth and increased demand, and thus net job growth. A commonly cited example is that of bank tellers: although ATMs replaced humans in the job of counting out cash for withdrawals, that made it cheaper to operate a bank branch, so the number of branches increased, leading to more bank employees overall. The nature of the work also changed, becoming less routine and requiring more advanced business skills. The net effect of automation seems to be in eliminating *tasks* rather than *jobs*.

The majority of commenters predict that the same will hold true with AI technology, at least in the short run. Gartner, McKinsey, Forbes, the World Economic Forum, and the Pew Research Center each released reports in 2018 predicting a net increase in jobs due to AI-driven automation. But some analysts think that this time around, things will be different. In 2019, IBM predicted that 120 million workers would need retraining due to automation by 2022, and Oxford Economics predicted that 20 million manufacturing jobs could be lost to automation by 2030.

Frey and Osborne (2017) survey 702 different occupations, and estimate that 47% of them are at risk of being automated, meaning that at least some of the tasks in the occupation can be performed by machine. For example, almost 3% of the workforce in the U.S. are vehicle drivers, and in some districts, as much as 15% of the male workforce are drivers. As we saw in Chapter 26, the task of driving is likely to be eliminated by driverless cars/trucks/buses/taxis.

It is important to distinguish between occupations and the tasks within those occupations. McKinsey estimates that only 5% of occupations are fully automatable, but that 60% of occupations can have about 30% of their tasks automated. For example, future truck drivers will spend less time holding the steering wheel and more time making sure that the goods are picked up and delivered properly; serving as customer service representatives and salespeople at either end of the journey; and perhaps managing convoys of, say, three robotic trucks. Replacing three drivers with one convoy manager implies a net loss in employment, but if transportation costs decrease, there will be more demand, which wins some of the jobs back—but perhaps not all of them. As another example, despite many advances in applying machine learning to the problem of medical imaging, radiologists have so far been augmented, not replaced, by these tools. Ultimately, there is a choice of how to make use of automation: do we want to focus on *cutting cost*, and thus see job loss as a positive; or do we want to focus on *improving quality*, making life better for the worker and the customer?

It is difficult to predict exact timelines for automation, but currently, and for the next few years, the emphasis is on automation of structured analytical tasks, such as reading x-ray images, customer relationship management (e.g., bots that automatically sort customer complaints and respond with suggested remedies), and **business process automation** that combines text documents and structured data to make business decisions and improve workflow. Over time, we will see more automation with physical robots, first in controlled warehouse environments, then in more uncertain environments, building to a significant portion of the marketplace by around 2030.

Business process automation

As populations in developed countries grow older, the ratio between workers and retirees changes. In 2015 there were less than 30 retirees per 100 workers; by 2050 there may be over 60 per 100 workers. Care for the elderly will be an increasingly important role, one that can partially be filled by AI. Moreover, if we want to maintain the current standard of living, it will also be necessary to make the remaining workers more productive; automation seems like the best opportunity to do that.

Even if automation has a multi-trillion-dollar net positive impact, there may still be problems due to the **pace of change**. Consider how change came to the farming industry: in 1900, over 40% of the U.S. workforce was in agriculture, but by 2000 that had fallen to 2%.[3] That

Pace of change

[3] In 2010, although only 2% of the U.S. workforce were actual farmers, over 25% of the population (80 million people) played the FARMVILLE game at least once.

is a huge disruption in the way we work, but it happened over a period of 100 years, and thus across generations, not in the lifetime of one worker.

Workers whose jobs are automated away this decade may have to retrain for a new profession within a few years—and then perhaps see their new profession automated and face yet another retraining period. Some may be happy to leave their old profession—we see that as the economy improves, trucking companies need to offer new incentives to hire enough drivers—but workers will be apprehensive about their new roles. To handle this, we as a society need to provide lifelong education, perhaps relying in part on online education driven by artificial intelligence (Martin, 2012). Bessen (2015) argues that workers will not see increases in income until they are trained to implement the new technologies, a process that takes time.

Technology tends to magnify **income inequality**. In an information economy marked by high-bandwidth global communication and zero-marginal-cost replication of intellectual property (what Frank and Cook (1996) call the "Winner-Take-All Society"), rewards tend to be concentrated. If farmer Ali is 10% better than farmer Bo, then Ali gets about 10% more income: Ali can charge slightly more for superior goods, but there is a limit on how much can be produced on the land, and how far it can be shipped. But if software app developer Cary is 10% better than Dana, it may be that Cary ends up with 99% of the global market. AI increases the pace of technological innovation and thus contributes to this overall trend, but AI also holds the promise of allowing us to take some time off and let our automated agents handle things for a while. Tim Ferriss (2007) recommends using automation and outsourcing to achieve a four-hour work week.

Income inequality

Before the industrial revolution, people worked as farmers or in other crafts, but didn't report to a **job** at a place of work and put in hours for an employer. But today, most adults in developed countries do just that, and the job serves three purposes: it fuels the production of the goods that society needs to flourish, it provides the income that the worker needs to live, and it gives the worker a sense of purpose, accomplishment, and social integration. With increasing automation, it may be that these three purposes become disaggregated—society's needs will be served in part by automation, and in the long run, individuals will get their sense of purpose from contributions other than work. Their income needs can be served by social policies that include a combination of free or inexpensive access to social services and education, portable health care, retirement, and education accounts, progressive tax rates, earned income tax credits, negative income tax, or universal basic income.

28.3.6 Robot rights

The question of robot consciousness, discussed in Section 28.2, is critical to the question of what rights, if any, robots should have. If they have no consciousness, no qualia, then few would argue that they deserve rights.

But if robots can feel pain, if they can dread death, if they are considered "persons," then the argument can be made (e.g., by Sparrow (2004)) that they have rights and deserve to have their rights recognized, just as slaves, women, and other historically oppressed groups have fought to have their rights recognized. The issue of robot personhood is often considered in fiction: from Pygmalion to Coppélia to Pinocchio to the movies *AI* and *Centennial Man*, we have the legend of a doll/robot coming to life and striving to be accepted as a human with human rights. In real life, Saudi Arabia made headlines by giving honorary citizenship to Sophia, a human-looking puppet capable of speaking preprogrammed lines.

If robots have rights, then they should not be enslaved, and there is a question of whether reprogramming them would be a kind of enslavement. Another ethical issue involves voting rights: a rich person could buy thousands of robots and program them to cast thousands of votes—should those votes count? If a robot clones itself, can they both vote? What is the boundary between ballot stuffing and exercising free will, and when does robotic voting violate the "one person, one vote" principle?

Ernie Davis argues for avoiding the dilemmas of robot consciousness by never building robots that could possibly be considered conscious. This argument was previously made by Joseph Weizenbaum in his book *Computer Power and Human Reason* (1976), and before that by Julien de La Mettrie in *L'Homme Machine* (1748). Robots are tools that we create, to do the tasks we direct them to do, and if we grant them personhood, we are just declining to take responsibility for the actions of our own property: "I'm not at fault for my self-driving car crash—the car did it itself."

This issue takes a different turn if we develop human–robot hybrids. Of course we already have humans enhanced by technology such as contact lenses, pacemakers, and artificial hips. But adding computational protheses may blur the lines between human and machine.

28.3.7 AI Safety

Almost any technology has the potential to cause harm in the wrong hands, but with AI and robotics, the hands might be operating on their own. Countless science fiction stories have warned about robots or cyborgs running amok. Early examples include Mary Shelley's *Frankenstein, or the Modern Prometheus* (1818) and Karel Čapek's play *R.U.R.* (1920), in which robots conquer the world. In movies, we have *The Terminator* (1984) and *The Matrix* (1999), which both feature robots trying to eliminate humans—the **robopocalypse** (Wilson, 2011). Perhaps robots are so often the villains because they represent the unknown, just like the witches and ghosts of tales from earlier eras. We can hope that a robot that is smart enough to figure out how to terminate the human race is also smart enough to figure out that that was not the intended utility function; but in building intelligent systems, we want to rely not just on hope, but on a design process with guarantees of safety.

Robopocalypse

It would be unethical to distribute an unsafe AI agent. We require our agents to avoid accidents, to be resistant to adversarial attacks and malicious abuse, and in general to cause benefits, not harms. That is especially true as AI agents are deployed in safety-critical applications, such as driving cars, controlling robots in dangerous factory or construction settings, and making life-or-death medical decisions.

Safety engineering

There is a long history of **safety engineering** in traditional engineering fields. We know how to build bridges, airplanes, spacecraft, and power plants that are designed up front to behave safely even when components of the system fail. The first technique is **failure modes and effect analysis (FMEA)**: analysts consider each component of the system, and imagine every possible way the component could go wrong (for example, what if this bolt were to snap?), drawing on past experience and on calculations based on the physical properties of the component. Then the analysts work forward to see what would result from the failure. If the result is severe (a section of the bridge could fall down) then the analysts alter the design to mitigate the failure. (With this additional cross-member, the bridge can survive the failure of any 5 bolts; with this backup server, the online service can survive a tsunami taking out the primary server.) The technique of **fault tree analysis (FTA)** is used to make these

Failure modes and effect analysis (FMEA)

Fault tree analysis (FTA)

determinations: analysts build an AND/OR tree of possible failures and assign probabilities to each root cause, allowing for calculations of overall failure probability. These techniques can and should be applied to all safety-critical engineered systems, including AI systems.

The field of **software engineering** is aimed at producing reliable software, but the emphasis has historically been on *correctness*, not *safety*. Correctness means that the software faithfully implements the specification. But safety goes beyond that to insist that the specification has considered any feasible failure modes, and is designed to degrade gracefully even in the face of unforeseen failures. For example, the software for a self-driving car wouldn't be considered safe unless it can handle unusual situations. For example, what if the power to the main computer dies? A safe system will have a backup computer with a separate power supply. What if a tire is punctured at high speed? A safe system will have tested for this, and will have software to correct for the resulting loss of control.

An agent designed as a utility maximizer, or as a goal achiever, can be unsafe if it has the wrong objective function. Suppose we give a robot the task of fetching a coffee from the kitchen. We might run into trouble with **unintended side effects**—the robot might rush to accomplish the goal, knocking over lamps and tables along the way. In testing, we might notice this kind of behavior and modify the utility function to penalize such damage, but it is difficult for the designers and testers to anticipate *all* possible side effects ahead of time.

Unintended side effect

One way to deal with this is to design a robot to have **low impact** (Armstrong and Levinstein, 2017): instead of just maximizing utility, maximize the utility minus a weighted summary of all changes to the state of the world. In this way, all other things being equal, the robot prefers not to change those things whose effect on utility is unknown; so it avoids knocking over the lamp not because it knows specifically that knocking the lamp will cause it to fall over and break, but because it knows in general that disruption might be bad. This can be seen as a version of the physician's creed "first, do no harm," or as an analog to **regularization** in machine learning: we want a policy that achieves goals, but we prefer policies that take smooth, low-impact actions to get there. The trick is how to measure impact. It is not acceptable to knock over a fragile lamp, but perfectly fine if the air molecules in the room are disturbed a little, or if some bacteria in the room are inadvertently killed. It is certainly not acceptable to harm pets and humans in the room. We need to make sure that the robot knows the differences between these cases (and many subtle cases in between) through a combination of explicit programming, machine learning over time, and rigorous testing.

Low impact

Utility functions can go wrong due to **externalities**, the word used by economists for factors that are outside of what is measured and paid for. The world suffers when greenhouse gases are considered as externalities—companies and countries are not penalized for producing them, and as a result everyone suffers. Ecologist Garrett Hardin (1968) called the exploitation of shared resources the **tragedy of the commons**. We can mitigate the tragedy by internalizing the externalities—making them part of the utility function, for example with a carbon tax—or by using the design principles that economist Elinor Ostrom identified as being used by local people throughout the world for centuries (work that won her the Nobel Prize in Economics in 2009):

- Clearly define the shared resource and who has access.
- Adapt to local conditions.
- Allow all parties to participate in decisions.

- Monitor the resource with accountable monitors.
- Sanctions, proportional to the severity of the violation.
- Easy conflict resolution procedures.
- Hierarchical control for large shared resources.

Victoria Krakovna (2018) has cataloged examples of AI agents that have gamed the system, figuring out how to maximize utility without actually solving the problem that their designers intended them to solve. To the designers this looks like cheating, but to the agents, they are just doing their job. Some agents took advantage of bugs in the simulation (such as floating point overflow bugs) to propose solutions that would not work once the bug was fixed. Several agents in video games discovered ways to crash or pause the game when they were about to lose, thus avoiding a penalty. And in a specification where crashing the game was penalized, one agent learned to use up just enough of the game's memory so that when it was the opponent's turn, it would run out of memory and crash the game. Finally, a genetic algorithm operating in a simulated world was supposed to evolve fast-moving creatures but in fact produced creatures that were enormously tall and moved fast by falling over.

Designers of agents should be aware of these kinds of specification failures and take steps to avoid them. To help them do that, Krakovna was part of the team that released the AI Safety Gridworlds environments (Leike *et al.*, 2017), which allows designers to test how well their agents perform.

Value alignment problem

The moral is that we need to be very careful in specifying what we want, because with utility maximizers we get what we actually asked for. The **value alignment problem** is the problem of making sure that what we ask for is what we really want; it is also known as the **King Midas problem**, as discussed on page 51. We run into trouble when a utility function fails to capture background societal norms about acceptable behavior. For example, a human who is hired to clean floors, when faced with a messy person who repeatedly tracks in dirt, knows that it is acceptable to politely ask the person to be more careful, but it is not acceptable to kidnap or incapacitate said person.

A robotic cleaner needs to know these things too, either through explicit programming or by learning from observation. Trying to write down all the rules so that the robot always does the right thing is almost certainly hopeless. We have been trying to write loophole-free tax laws for several thousand years without success. Better to make the robot *want* to pay taxes, so to speak, than to try to make rules to force it to do so when it really wants to do something else. A sufficiently intelligent robot will find a way to do something else.

Robots can learn to conform better with human preferences by observing human behavior. This is clearly related to the notion of apprenticeship learning (Section 23.6). The robot may learn a policy that directly suggests what actions to take in what situations; this is often a straightforward supervised learning problem if the environment is observable. For example, a robot can watch a human playing chess: each state–action pair is an example for the learning process. Unfortunately, this form of **imitation learning** means that the robot will repeat human mistakes. Instead, the robot can apply **inverse reinforcement learning** to discover the utility function that the humans must be operating under. Watching even terrible chess players is probably enough for the robot to learn the objective of the game. Given just this information, the robot can then go on to exceed human performance—as, for example, ALPHAZERO did in chess—by computing optimal or near-optimal policies from the objec-

tive. This approach works not just in board games, but in real-world physical tasks such as helicopter aerobatics (Coates *et al.*, 2009).

In more complex settings involving, for example, social interactions with humans, it is very unlikely that the robot will converge to exact and correct knowledge of each human's individual preferences. (After all, many humans never quite learn what makes other humans tick, despite a lifetime of experience, and many of us are unsure of our own preferences too.) It will be necessary, therefore, for machines to function appropriately when it is uncertain about human preferences. In Chapter 17, we introduced **assistance games**, which capture exactly this situation. Solutions to assistance games include acting cautiously, so as not to disturb aspects of the world that the human might care about, and asking questions. For example, the robot could ask whether turning the oceans into sulphuric acid is an acceptable solution to global warming before it puts the plan into effect.

In dealing with humans, a robot solving an assistance game must accommodate human imperfections. If the robot asks permission, the human may give it, not foreseeing that the robot's proposal is in fact catastrophic in the long term. Moreover, humans do not have complete introspective access to their true utility function, and they don't always act in a way that is compatible with it. Humans sometimes lie or cheat, or do things they know are wrong. They sometimes take self-destructive actions like overeating or abusing drugs. AI systems need not learn to adopt these problematic tendencies, but they must understand that they exist when interpreting human behavior to get at the underlying human preferences.

Despite this toolbox of safeguards, there is a fear, expressed by prominent technologists such as Bill Gates and Elon Musk and scientists such as Stephen Hawking and Martin Rees, that AI could evolve out of control. They warn that we have no experience controlling powerful nonhuman entities with super-human capabilities. However, that's not quite true; we have centuries of experience with nations and corporations; non-human entities that aggregate the power of thousands or millions of people. Our record of controlling these entities is not very encouraging: nations produce periodic convulsions called wars that kill tens of millions of human beings, and corporations are partly responsible for global warming and our inability to confront it.

AI systems may present much greater problems than nations and corporations because of their potential to self-improve at a rapid pace, as considered by I. J. Good (1965b):

Ultraintelligent machine

> Let an **ultraintelligent machine** be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.

Technological singularity

Good's "intelligence explosion" has also been called the **technological singularity** by mathematics professor and science fiction author Vernor Vinge, who wrote in 1993: "Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended." In 2017, inventor and futurist Ray Kurzweil predicted the singularity would appear by 2045, which means it got 2 years closer in 24 years. (At that rate, only 336 years to go!) Vinge and Kurzweil correctly note that technological progress on many measures is growing exponentially at present.

It is, however, quite a leap to extrapolate all the way from the rapidly decreasing cost of computation to a singularity. So far, every technology has followed an S-shaped curve, where the exponential growth eventually tapers off. Sometimes new technologies step in when the old ones plateau, but sometimes it is not possible to keep the growth going, for technical, political, or sociological reasons. For example, the technology of flight advanced dramatically from the Wright brothers' flight in 1903 to the moon landing in 1969, but has had no breakthroughs of comparable magnitude since then.

Another obstacle in the way of ultraintelligent machines taking over the world is the world. More specifically, some kinds of progress require not just thinking but acting in the physical world. (Kevin Kelly calls the overemphasis on pure intelligence **thinkism**.) An ultraintelligent machine tasked with creating a grand unified theory of physics might be capable of cleverly manipulating equations a billion times faster than Einstein, but to make any real progress, it would still need to raise millions of dollars to build a more powerful supercollider and run physical experiments over the course of months or years. Only then could it start analyzing the data and theorizing. Depending on how the data turn out, the next step might require raising additional billions of dollars for an interstellar probe mission that would take centuries to complete. The "ultraintelligent thinking" part of this whole process might actually be the least important part. As another example, an ultraintelligent machine tasked with bringing peace to the Middle East might just end up getting 1000 times more frustrated than a human envoy. As yet, we don't know how many of the big problems are like mathematics and how many are like the Middle East.

Thinkism

While some people fear the singularity, others relish it. The **transhumanism** social movement looks forward to a future in which humans are merged with—or replaced by—robotic and biotech inventions. Ray Kurzweil writes in *The Singularity is Near* (2005):

Transhumanism

> The Singularity will allow us to transcend these limitations of our biological bodies and brain. We will gain power over our fates. . . . We will be able to live as long as we want . . . We will fully understand human thinking and will vastly extend and expand its reach. By the end of this century, the nonbiological portion of our intelligence will be trillions of trillions of times more powerful than unaided human intelligence.

Similarly, when asked whether robots will inherit the Earth, Marvin Minsky said "yes, but they will be our children." These possibilities present a challenge for most moral theorists, who take the preservation of human life and the human species to be a good thing. Kurzweil also notes the potential dangers, writing "But the Singularity will also amplify the ability to act on our destructive inclinations, so its full story has not yet been written." We humans would do well to make sure that any intelligent machine we design today that might evolve into an ultraintelligent machine will do so in a way that ends up treating us well. As Eric Brynjolfsson puts it, "The future is not preordained by machines. It's created by humans."

Summary

This chapter has addressed the following issues:

- Philosophers use the term **weak AI** for the hypothesis that machines could possibly behave intelligently, and **strong AI** for the hypothesis that such machines would count as having actual minds (as opposed to simulated minds).

- Alan Turing rejected the question "Can machines think?" and replaced it with a behavioral test. He anticipated many objections to the possibility of thinking machines. Few AI researchers pay attention to the Turing test, preferring to concentrate on their systems' performance on practical tasks, rather than the ability to imitate humans.
- Consciousness remains a mystery.
- AI is a powerful technology, and as such it poses potential dangers, through lethal autonomous weapons, security and privacy breaches, unintended side effects, unintentional errors, and malignant misuse. Those who work with AI technology have an ethical imperative to responsibly reduce those dangers.
- AI systems must be able to demonstrate they are fair, trustworthy, and transparent.
- There are multiple aspects of fairness, and it is impossible to maximize all of them at once. So a first step is to decide what counts as fair.
- Automation is already changing the way people work. As a society, we will have to deal with these changes.

Bibliographical and Historical Notes

Weak AI: When Alan Turing (1950) proposed the possibility of AI, he also posed many of the key philosophical questions, and provided possible replies. But various philosophers had raised similar issues long before AI was invented. Maurice Merleau-Ponty's *Phenomenology of Perception* (1945) stressed the importance of the body and the subjective interpretation of reality afforded by our senses, and Martin Heidegger's *Being and Time* (1927) asked what it means to actually be an agent. In the computer age, Alva Noe (2009) and Andy Clark (2015) propose that our brains form a rather minimal representation of the world, use the world itself on a just-in-time basis to maintain the illusion of a detailed internal model, and use props in the world (such as paper and pencil as well as computers) to increase the capabilities of the mind. Pfeifer *et al.* (2006) and Lakoff and Johnson (1999) present arguments for how the body helps shape cognition. Speaking of bodies, Levy (2008), Danaher and McArthur (2017), and Devlin (2018) address the issue of robot sex.

Strong AI: René Descartes is known for his dualistic view of the human mind, but ironically his historical influence was toward mechanism and physicalism. He explicitly conceived of animals as automata, and he anticipated the Turing test, writing "it is not conceivable [that a machine] should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as even the dullest of men can do" (Descartes, 1637). Descartes's spirited defense of the animals-as-automata viewpoint actually had the effect of making it easier to conceive of humans as automata as well, even though he himself did not take this step. The book *L'Homme Machine* (La Mettrie, 1748) did explicitly argue that humans are automata. As far back as Homer (circa 700 BCE), the Greek legends envisioned automata such as the bronze giant Talos and considered the issue of *biotechne*, or life through craft (Mayor, 2018).

The **Turing test** (Turing, 1950) has been debated (Shieber, 2004), anthologized (Epstein *et al.*, 2008), and criticized (Shieber, 1994; Ford and Hayes, 1995). Bringsjord (2008) gives advice for a Turing test judge, and Christian (2011) for a human contestant. The annual Loebner Prize competition is the longest-running Turing test-like contest; Steve Worswick's

MITSUKU won four in a row from 2016 to 2019. The **Chinese room** has been debated endlessly (Searle, 1980; Chalmers, 1992; Preston and Bishop, 2002). Hernández-Orallo (2016) gives an overview of approaches to measuring AI progress, and Chollet (2019) proposes a measure of intelligence based on skill-acquisition efficiency.

Consciousness remains a vexing problem for philosophers, neuroscientists, and anyone who has pondered their own existence. Block (2009), Churchland (2013) and Dehaene (2014) provide overviews of the major theories. Crick and Koch (2003) add their expertise in biology and neuroscience to the debate, and Gazzaniga (2018) shows what can be learned from studying brain disabilities in hospital cases. Koch (2019) gives a theory of consciousness—"intelligence is about doing while experience is about being"—that includes most animals, but not computers. Giulio Tononi and his colleagues propose **integrated information theory** (Oizumi *et al.*, 2014). Damasio (1999) has a theory based on three levels: emotion, feeling, and feeling a feeling. Bryson (2012) shows the value of conscious attention for the process of learning action selection.

The philosophical literature on minds, brains, and related topics is large and jargon-filled. The *Encyclopedia of Philosophy* (Edwards, 1967) is an impressively authoritative and very useful navigation aid. *The Cambridge Dictionary of Philosophy* (Audi, 1999) is shorter and more accessible, and the online *Stanford Encyclopedia of Philosophy* offers many excellent articles and up-to-date references. The *MIT Encyclopedia of Cognitive Science* (Wilson and Keil, 1999) covers the philosophy, biology, and psychology of mind. There are multiple introductions to the philosophical "AI question" (Haugeland, 1985; Boden, 1990; Copeland, 1993; McCorduck, 2004; Minsky, 2007). *The Behavioral and Brain Sciences*, abbreviated *BBS*, is a major journal devoted to philosophical and scientific debates about AI and neuroscience.

Science fiction writer Isaac Asimov (1942, 1950) was one of the first to address the issue of robot ethics, with his **laws of robotics**:

0. A robot may not harm humanity, or through inaction, allow humanity to come to harm.
1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

At first glance, these laws seem reasonable. But the trick is how to implement them. Should a robot allow a human to cross the street, or eat junk food, if the human might conceivably come to harm? In Asimov's story *Runaround* (1942), humans need to debug a robot that is found wandering in a circle, acting "drunk." They work out that the circle defines the locus of points that balance the second law (the robot was ordered to fetch some selenium at the center of the circle) with the third law (there is a danger there that threatens the robot's existence).[4] This suggests that the laws are not logical absolutes, but rather are weighed against each other, with a higher weight for the earlier laws. As this was 1942, before the emergence of

[4] Science fiction writers are in broad agreement that robots are very bad at resolving contradictions. In *2001*, the HAL 9000 computer becomes homicidal due to a conflict in its orders, and in the *Star Trek* episode "I, Mudd," Captain Kirk tells an enemy robot that "Everything Harry tells you is a lie," and Harry says "I am lying." At this, smoke comes out of the robot's head and it shuts down.

digital computers, Asimov was probably thinking of an architecture based on control theory via analog computing.

Weld and Etzioni (1994) analyze Asimov's laws and suggest some ways to modify the planning techniques of Chapter 11 to generate plans that do no harm. Asimov has considered many of the ethical issues around technology; in his 1958 story *The Feeling of Power* he tackles the issue of automation leading to a lapse of human skill—a technician rediscovers the lost art of multiplication—as well as the dilemma of what to do when the rediscovery is applied to warfare.

Norbert Wiener's book *God & Golem, Inc.* (1964) correctly predicted that computers would achieve expert-level performance at games and other tasks, and that specifying what it is that we want would prove to be difficult. Wiener writes:

> While it is always possible to ask for something other than we really want, this possibility is most serious when the process by which we are to obtain our wish is indirect, and the degree to which we have obtained our wish is not clear until the very end. Usually we realize our wishes, insofar as we do actually realize them, by a feedback process, in which we compare the degree of attainment of intermediate goals with our anticipation of them. In this process, the feedback goes through us, and we can turn back before it is too late. If the feedback is built into a machine that cannot be inspected until the final goal is attained, the possibilities for catastrophe are greatly increased. I should very much hate to ride on the first trial of an automobile regulated by photoelectric feedback devices, unless there were somewhere a handle by which I could take over control if I found myself driving smack into a tree.

We summarized **codes of ethics** in the chapter, but the list of organizations that have issued sets of principles is growing rapidly, and now includes Apple, DeepMind, Facebook, Google, IBM, Microsoft, the Organisation for Economic Co-operation and Development (OECD), the United Nations Educational, Scientific and Cultural Organization (UNESCO), the U.S. Office of Science and Technology Policy the Beijing Academy of Artificial Intelligence (BAAI), the Institute of Electrical and Electronics Engineers (IEEE), the Association of Computing Machinery (ACM), the World Economic Forum, the Group of Twenty (G20), OpenAI, the Machine Intelligence Research Institute (MIRI), AI4People, the Centre for the Study of Existential Risk, the Center for Human-Compatible AI, the Center for Humane Technology, the Partnership on AI, the AI Now Institute, the Future of Life Institute, the Future of Humanity Institute, the European Union, and at least 42 national governments. We have the handbook on the *Ethics of Computing* (Berleur and Brunnstein, 2001) and introductions to the topic of AI ethics in book (Boddington, 2017) and survey (Etzioni and Etzioni, 2017a) form. The *Journal of Artificial Intelligence and Law* and *AI and Society* cover ethical issues. We'll now look at some of the individual issues.

Lethal autonomous weapons: P. W. Singer's *Wired for War* (2009) raised ethical, legal, and technical issues around robots on the battlefield. Paul Scharre's *Army of None* (2018), written by one of the authors of current US policy on autonomous weapons, offers a balanced and authoritative view. Etzioni and Etzioni (2017b) address the question of whether artificial intelligence should be regulated; they recommend a pause in the development of lethal autonomous weapons, and an international discussion on the subject of regulation.

Privacy: Latanya Sweeney (Sweeney, 2002b) presents the k-anonymity model and the idea of generalizing fields (Sweeney, 2002a). Achieving k-anonymity with minimal loss of data is an NP-hard problem, but Bayardo and Agrawal (2005) give an approximation algorithm. Cynthia Dwork (2008) describes differential privacy, and in subsequent work gives practical examples of clever ways to apply differential privacy to get better results than the naive approach (Dwork *et al.*, 2014). Guo *et al.* (2019) describe a process for certified data removal: if you train a model on some data, and then there is a request to delete some of the data, this extension of differential privacy lets you modify the model and prove that it does not make use of the deleted data. Ji *et al.* (2014) gives a review of the field of privacy. Etzioni (2004) argues for a balancing of privacy and security; individual rights and community. Fung *et al.* (2018), Bagdasaryan *et al.* (2018) discuss the various attacks on federated learning protocols. Narayanan *et al.* (2011) describe how they were able to de-anonymize the obfuscated connection graph from the 2011 Social Network Challenge by crawling the site where the data was obtained (Flickr), and matching nodes with unusually high in-degree or out-degree between the provided data and the crawled data. This allowed them to gain additional information to win the challenge, and it also allowed them to uncover the true identity of nodes in the data. Tools for user privacy are becoming available; for example, TensorFlow provides modules for federated learning and privacy (McMahan and Andrew, 2018).

Fairness: Cathy O'Neil's book *Weapons of Math Destruction* (2017) describes how various black box machine learning models influence our lives, often in unfair ways. She calls on model builders to take responsibility for fairness, and for policy makers to impose appropriate regulation. Dwork *et al.* (2012) showed the flaws with the simplistic "fairness through unawareness" approach. Bellamy *et al.* (2018) present a toolkit for mitigating bias in machine learning systems. Tramèr *et al.* (2016) show how an adversary can "steal" a machine learning model by making queries against an API, Hardt *et al.* (2017) describe equal opportunity as a metric for fairness. Chouldechova and Roth (2018) give an overview of the frontiers of fairness, and Verma and Rubin (2018) give an exhaustive survey of fairness definitions.

Kleinberg *et al.* (2016) show that, in general, an algorithm cannot be both well-calibrated and equal opportunity. Berk *et al.* (2017) give some additional definitions of types of fairness, and again conclude that it is impossible to satisfy all aspects at once. Beutel *et al.* (2019) give advice for how to put fairness metrics into practice.

Dressel and Farid (2018) report on the COMPAS recidivism scoring model. Christin *et al.* (2015) and Eckhouse *et al.* (2019) discuss the use of predictive algorithms in the legal system. Corbett-Davies *et al.* (2017) show that that there is a tension between ensuring fairness and optimizing public safety, and Corbett-Davies and Goel (2018) discuss the differences between fairness frameworks. Chouldechova (2017) advocates for fair impact: all classes should have the same expected utility. Liu *et al.* (2018a) advocate for a long-term measure of impact, pointing out that, for example, if we change the decision point for approving a loan in order to be more fair in the short run, this could have negative effect in the long run on people who end up defaulting on a loan and thus have their credit score reduced.

Since 2014 there has been an annual conference on Fairness, Accountability, and Transparency in Machine Learning. Mehrabi *et al.* (2019) give a comprehensive survey of bias and fairness in machine learning, cataloging 23 kinds of bias and 10 definitions of fairness.

Trust: Explainable AI was an important topic going back to the early days of expert systems (Neches *et al.*, 1985), and has been making a resurgence in recent years (Biran and

Cotton, 2017; Miller *et al.*, 2017; Kim, 2018). Barreno *et al.* (2010) give a taxonomy of the types of security attacks that can be made against a machine learning system, and Tygar (2011) surveys adversarial machine learning. Researchers at IBM have a proposal for gaining trust in AI systems through declarations of conformity (Hind *et al.*, 2018). DARPA requires explainable decisions for its battlefield systems, and has issued a call for research in the area (Gunning, 2016).

AI safety: The book *Artificial Intelligence Safety and Security* (Yampolskiy, 2018) collects essays on AI safety, both recent and classic, going back to Bill Joy's *Why the Future Doesn't Need Us* (Joy, 2000). The "King Midas problem" was anticipated by Marvin Minsky, who once suggested that an AI program designed to solve the Riemann Hypothesis might end up taking over all the resources of Earth to build more powerful supercomputers. Similarly, Omohundro (2008) foresees a chess program that hijacks resources, and Bostrom (2014) describes the runaway paper clip factory that takes over the world. Yudkowsky (2008) goes into more detail about how to design a **Friendly AI**. Amodei *et al.* (2016) present five practical safety problems for AI systems.

Omohundro (2008) describes the *Basic AI Drives* and concludes, "Social structures which cause individuals to bear the cost of their negative externalities would go a long way toward ensuring a stable and positive future." Elinor Ostrom's *Governing the Commons* (1990) describes practices for dealing with externalities by traditional cultures. Ostrom has also applied this approach to the idea of knowledge as a commons (Hess and Ostrom, 2007).

Ray Kurzweil (2005) proclaimed *The Singularity is Near*, and a decade later Murray Shanahan (2015) gave an update on the topic. Microsoft cofounder Paul Allen countered with *The Singularity isn't Near* (2011). He didn't dispute the possibility of ultraintelligent machines; he just thought it would take more than a century to get there. Rod Brooks is a frequent critic of singularitarianism; he points out that technologies often take longer than predicted to mature, that we are prone to magical thinking, and that exponentials don't last forever (Brooks, 2017).

On the other hand, for every optimistic singularitarian there is a pessimist who fears new technology. The Web site `pessimists.co` shows that this has been true throughout history: for example, in the 1890s people were concerned that the elevator would inevitably cause nausea, that the telegraph would lead to loss of privacy and moral corruption, that the subway would release dangerous underground air and disturb the dead, and that the bicycle—especially the idea of a woman riding one—was the work of the devil.

Hans Moravec (2000) introduces some of the ideas of transhumanism, and Bostrom (2005) gives an updated history. Good's ultraintelligent machine idea was foreseen a hundred years earlier in Samuel Butler's *Darwin Among the Machines* (1863). Written four years after the publication of Charles Darwin's *On the Origins of Species* and at a time when the most sophisticated machines were steam engines, Butler's article envisioned "the ultimate development of mechanical consciousness" by natural selection. The theme was reiterated by George Dyson (1998) in a book of the same title, and was referenced by Alan Turing, who wrote in 1951 "At some stage therefore we should have to expect the machines to take control in the way that is mentioned in Samuel Butler's *Erewhon*" (Turing, 1996).

Robot rights: A book edited by Yorick Wilks (2010) gives different perspectives on how we should deal with artificial companions, ranging from Joanna Bryson's view that robots should serve us as tools, not as citizens, to Sherry Turkle's observation that we already per-

sonify our computers and other tools, and are quite willing to blur the boundaries between machines and life. Wilks also contributed a recent update on his views (Wilks, 2019). The philosopher David Gunkel's book *Robot Rights* (2018) considers four possibilities: *can* robots have rights or not, and *should* they or not? The American Society for the Prevention of Cruelty to Robots (ASPCR) proclaims that "The ASPCR is, and will continue to be, exactly as serious as robots are sentient."

The future of work: In 1888, Edward Bellamy published the best-seller *Looking Backward*, which predicted that by the year 2000, technological advances would led to a utopia where equality is achieved and people work short hours and retire early. Soon after, E. M. Forster took the dystopian view in *The Machine Stops* (1909), in which a benevolent machine takes over the running of a society; things fall apart when the machine inevitably fails. Norbert Wiener's prescient book *The Human Use of Human Beings* (1950) argues for the benefits of automation in freeing people from drudgery while offering more creative work, but also discusses several dangers that we recognize as problems today, particularly the problem of value alignment.

The book *Disrupting Unemployment* (Nordfors *et al.*, 2018) discuss some of the ways that work is changing, opening opportunities for new careers. Erik Brynjolfsson and Andrew McAfee address these themes and more in their books *Race Against the Machine* (2011) and *The Second Machine Age* (2014). Ford (2015) describes the challenges of increasing automation, and West (2018) provides recommendations to mitigate the problems, while MIT's Thomas Malone (2004) shows that many of the same issues were apparent a decade earlier, but at that time were attributed to worldwide communication networks, not to automation.