CHAPTER 27

PHILOSOPHY AND ETHICS OF AI

In which we consider the big questions.

Philosophers have been asking big questions for centuries: 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? What are the ethical implications of intelligent machines?

27.1 The Limits of AI

Weak AI

Strong AI

In 1980, philosopher John Searle introduced a distinction between **weak AI**—the idea that machines could use clever "tricks" to 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 as well as a human.

Critics who objected to the very possibility of intelligent machines now look like Simon Newcomb, who in 1903 wrote "aerial flight is one of the great class of problems with which man can never cope." That same year, the Wright brothers proved him wrong, and every month new advances prove the weak AI critics wrong. However, there may well be *some* 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.

27.1.1 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 principal proponent of this view was the philosopher 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). In a similar vein, 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 position they criticize came to be called "Good Old-Fashioned AI," or GOFAI (Haugeland, 1985). 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 21, 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—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." That means, Clark says, that 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.

27.1.2 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 random access memory can access that memory, 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, and the 2013 movie *Her* explores that theme. 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 done things that are "really new," making small but 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.

[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).

27.1.3 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) that repeat the claim with some fresh twists, such as the hypothesis that humans are different because their brains operate by quantum gravity—a theory that makes several 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 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 a reasonable approximation, but 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 (but has other sources of incompleteness: a finite system with n bits of memory can represent at most 2^n proofs, so there must be true proofs that it cannot represent). Lucas assumes that humans can "change their minds" while computers are fixed and cannot, but that is also false—a computer with effectors connected to the world can change its hardware by, say, mining bitcoin and using it to buy a memory upgrade, or can change its software by recompiling pieces of itself, or can change its decision-making process just by updating its internal representations.

27.1.4 Measuring AI

Turing Test

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 is for 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 was not the exact details of the test, but rather the idea of measuring intelligence by performance on some kind of behavioral test, 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 cover up for any grammatical errors. Perhaps the Turing test is really a test of human gullibility. To date, no well-trained judge has been fooled for an extended period of time (Aaronson, 2014).

Turing test competitions have led to better chatbots, but have not attracted much interest within the AI community. Instead, an AI researchers who wants to participate in a competition is more likely to concentrate on a task such as playing chess or Go or Starcraft II against a human champion, 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.

27.2 Can Machines Really Think?

Can submarines swim?

Some philosophers have claimed that a machine that acts intelligently would not be *actually* thinking, but would be only a *simulation* of thinking. But most AI researchers don't care about 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 position known as 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 it is more important for the machine to have a humanoid appearance and voice than for it to actually be intelligent.

Polite convention

27.2.1 The 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 answers. Searle then argues: the person in the room 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.

Chinese room

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 same 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 think. How they can remains a mystery.

Biological naturalism

27.2.2 Consciousness and qualia

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 a programmed response, no different from "Error 404: not found." There is a similar question for animals: pet owners are usually quite certain that their dog or cat has consciousness, but not all scientists agree. Crickets change their behavior based on temperature, but most people would say that they do not

Consciousness

Qualia

experience the *feeling* of being warm or cold. One reason that the problem of consciousness is hard is that it remains ill-defined.

Philosophers have debated consciousness for centuries, but recently they 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 consciousness.

27.3 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 is a tool that can help save lives through medical discoveries and improved treatments, better prediction of extreme weather events, and with self-driving cars and driver assistance systems that prevent accidents. There are also many opportunities to improve lives. Microsoft's AI for Humanitarian Action program applies AI to disaster recovery, 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, monitoring of ground water, air, mines, pesticides, and 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).

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

Negative side effects

the threat of global destruction; the internal combustion engine brought air pollution, global warming, and the paving of paradise. In a sense, automobiles are robots that we have allowed to conquer the world by making them indispensable. 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, and in developed countries, some of that wealth will be widely distributed, but much will be concentrated in the hands of a few, increasing income inequality. This can be disruptive to a well-functioning society. In developing countries without widespread Internet access, income inequality will widen further. Our ethical and governance decisions will dictate the level of inequality.

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

27.3.1 Lethal autonomous weapons

Autonomous weapons have a long history. One could say that the crossbow, invented about 2500 years ago, was the first technology to enable human-initiated action to result in semi-reliable killing at a distance. More autonomous weapons soon followed: land mines have been used since the 17th century, guided missiles since the 1940s, and auto-firing radar-controlled guns have been used to defend naval ships since the 1970s. AI systems are now commonplace on the battlefield; the U.S. military deployed over 5,000 autonomous aircraft and 12,000 autonomous ground vehicles in the 2003 Iraq war and requested $3.7 billion to spend on autonomous drones in 2020.

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." There is a reluctance to completely remove the human from the control loop, for reasons of morality and reliability. Leaders remember the lesson of 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.

Third offset

The DOD has long relied on technology to provide an advantage or "offset" over potential opponents. The first offset involved nuclear deterrence, the second stressed stealth, intelligence gathering, and reconnaissance, and the **third offset**, currently under development, relies on "robotics, autonomous systems, miniaturization, big data, and advanced manufacturing," according to former Secretary of Defense Chuck Hagel (2014).

One way to look at it is that military robots are like medieval armor taken to its logical extreme: it is considered moral for a soldier to wear a helmet when being attacked by axe-wielding enemies, and a teleoperated robot is something like a very safe form of armor. Autonomous weapons are force multipliers, which means that when they are employed, fewer soldiers will be needed to accomplish a mission, thus fewer will risk injury, traumatic stress, or death.

Teleoperated military drones can accomplish their goals without putting a human pilot at risk, but they have killed thousands of innocent civilians (Benjamin, 2013) when the teleoperators make mistakes—in some cases the same kind of mistakes that a high-altitude pilot might make, and in some cases different. Taking the teleoperator out of the loop changes the equation, but it is not yet clear exactly how. Military operations are inherently uncertain—the fog of war—and adversarial attacks are expected. Tactically, that means that machine learning systems that operated flawlessly in training may perform poorly when deployed. A teleoperated system can be commanded to stop when it appears to have gone awry; it is not always possible to stop a fully autonomous system. Strategically, the availability of powerful robots might lower the barriers to war, and those who are attacked by autonomous weapons may feel additional moral outrage, making them less likely to accept a peaceful resolution without further revenge.

Besides national armies, individuals or small groups also pose a risk of attacks. Autonomous technology including robots, drones, and cyberwarfare actions can amplify the destructive power of small groups, and shield them from reprisal, because their victims may not know who launched the attack. In general, weapons are destabilizing when they provide a large advantage to attackers over defenders. For example, the crossbow is a lethal technology, but it is not destabilizing, because once both sides have it, we can expect that roughly equal numbers of people on each side will be killed. Thus, there is no incentive to rush to war. Fleets of autonomous drones are potentially destabilizing, because they can be programmed to kill in large numbers without fear of reprisal. When destabilizing threats like this are identified, one response is to initiate regulations and treaties to mitigate their deployment; another response is to invest in research into defensive systems to repel such attacks.

It is possible that a fully autonomous drone or robot, making decisions locally, could produce better, more humane decisions than a teleoperated system, or a soldier on the battlefield. Autonomous systems will not succumb to fatigue, frustration, hysteria, fear, anger, or revenge, and will not forget the international humanitarian law that governs warfare. They will not fear for their lives and "shoot first, ask questions later" (Arkin, 2015). Certainly a fully autonomous weapons system could offer better defense by responding to attacks faster than a human; that is why ships have auto-fire radar-controlled guns. And for those landmines that have not been eliminated by the 1997 Ottawa Treaty, wouldn't it be better if they knew the difference between an innocent civilian and an enemy soldier?

On the other hand, robotic weapon systems might operate unpredictably when partially damaged, and might turn rogue under a cyberattack. Arms dealers might represent the systems as more capable than they actually are, leading them to be deployed in situations where they would act improperly.

In light of this uncertainty, the United Nations Group of Governmental Experts asserted in 2018 that "human responsibility for decisions on the use of weapons systems must be retained, since accountability cannot be transferred to machines." 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." Others agree with these calls (Sharkey, 2010), but Australia, Israel, Russia, South Korea, and the United States opposed a treaty to ban autonomous weapons. Some individuals and entire companies have pledged not to work on autonomous weapons, and the Future of Life Institute organized an open letter signed by 4,000 AI researchers[2] and 22,000 others calling for a ban on autonomous weapons. Whereas countries understandably do not want to fall behind their rivals in technological readiness, a runaway AI arms race would benefit no one (except the arms dealers), so it makes sense to have a well-structured international convention on the use of autonomous weapons, as we have for chemical and biological weapons.

Dual use

One difficulty is dealing with **dual use** technologies, those that have peaceful as well as lethal uses. For example, chlorine gas has peaceful purposes and is not listed as a chemical weapon; that has made it easier for regimes to use the gas in bombs. With artificial intelligence, almost all of the technology is dual use. Once we have a high-functioning autonomous drone or robot for general use, it is easy to turn it into a weapon simply by attaching an explosive and commanding it to seek out a target. Dealing with this will require carefully designed treaties, combined with the kind of regulation and police work that is done to distinguish between legitimate purchases of ammonium nitrate fertilizer by farmers and suspect purchases by potential bombers.

27.3.2 Surveillance, security, and privacy

Surveillance

Technology has often been invoked for purposes of **surveillance**. In 1843, Jeremy Bentham (the pioneer of utilitarianism) proposed a breakthrough in architectural technology: the **panopticon**, a circular-shaped prison in which a central guard could observe all inmates at the same time. In 1976, Joseph Weizenbaum warned that 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. There are about 2 million **surveillance cameras** in the U.K., and 20 million in China, which currently has 100% of Beijing under surveillance, and plans to cover the rest of the country. China and other countries have begun exporting surveillance technology to low-tech countries, some with reputations for mistreating their citizens citizens and disproportionately targeting marginalized communities. AI workers should be clear on what uses of surveillance are compatible with human rights, and decline to work

Surveillance cameras

[2] Including the two authors of this book.

on applications that are incompatible.

Cybersecurity

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 commandering 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 to do A/B testing on phishing attempts. 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.

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 (HIPPA) 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.

De-identification

Netflix Prize

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**: sharing health records, after 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).

K-anonymity

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 with age 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.

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, but no response would be given for other zip codes.

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

Differential privacy

$$\text{abs}(\log P(Q(D)=y) - \log P(Q(D+r)=y)) \leq \varepsilon\,.$$

In other words, whether any one individual 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 modern 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** (Konečný *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. Communication between the central server and the distributed phones can be done when the phones are plugged in and are on wifi, so as not to consume battery or bandwidth resources.

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

Secure aggregation

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.

Another issue is whether citizens should have the right to know when they are interacting with an AI system. 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 with another agent." 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..." In the future, lawmakers and citizens may become less concerned with this issue, just as the original red flag law was repealed once automobile speeds exceeded walking speed.

27.3.3 Fairness and bias

Societal 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:

- **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 discover latent variables (such as race) given other variables (such as zipcode), so this approach by itself does not work.
- **Equal outcome**: The idea that each demographic class gets the same results; they have **demographic parity**. For example, suppose we have to decide if we should approve loan applications; the goal is to approve those applicants who will pay back the loan

Demographic parity

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. But this approach has two flaws. First, it 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. Second, suppose we had a perfect oracle that could predict with 100% accuracy who would default. If it happened that the percentages of default were different for males and females, we could not use this oracle and maintain demographic parity; we would be forced to make worse decisions.

- **Equal opportunity**: Of the people who can pay back the loan, the same percentage of males and females get approved. If we had a perfect oracle, we could use it to achieve equal opportunity, and lacking an oracle we can retrospectively measure how close we come to equal opportunity. This approach is also called "balance."
- **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.

As a case study, COMPAS is a commercial system for recidivism (re-offense) scoring that assigns to a defendant in a criminal case a **risk score**, which is then used by judges to help make certain 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. 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. But 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.

Well-calibrated

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. If we give value to those fears 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 it is difficult or impossible to acquire unbiased ground truth data. We want to predict whether a defendant will commit a crime. But the data we have does not say who has *committed* a crime—all we know is who has been *arrested* for a crime. If the arresting officers are 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 will appear in the data, so if the judges making the release decisions are biased, the data may be biased. (That source of bias could be overcome by doing a randomized controlled trial: randomly assigning defendants to the "release" or "not release" groups, and observing how they behave. But society has decided this is not a viable approach).

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 education, 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 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 cause bias. 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. Sample size also comes into play in reinforcement learning models. We always have an exploration/exploitation tradeoff, but a minority class has fewer members, and thus each member is more likely to pay a higher price for exploration.

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? A first idea is to understand the limits of the data you are using. It has been suggested (Gebru *et al.*, 2018; Hind *et al.*, 2018) that data sets 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 whether their intended use is feasible or not. 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 sheets

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) provides 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. 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 white dress, but harder to recognize traditional African and Indian garb.

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.

27.3.4 Trust

Trust

It is one challenge to make an AI system 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 process of **verification and validation** (V&V). Validation means that the product's specifications are appropriate to the user's needs. Verification means that the product satisfies the specifications. We have an elaborate V&V methodology for engineering in general, and for traditional software development done by human coders; some 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 had apprehension and fear about electric power. Their 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 nowhere near this level of clarity. There is ongoing debate about what kind of certification is necessary, and to what extent it should be done by the government, or by independent certifiers such as UL, or through self-regulation by the product companies.

Transparency

Explainable AI

Another aspect of trust is **transparency**: consumers want to know what is going on inside a system, and that nothing is hidden from them. When an AI system turns you down for a loan, you deserve an explanation (and in Europe, the GDPR enforces this for you). If the system can explain itself, we call it **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. There is a potential that explanations from machine learning systems could be better than explanations from humans, because machines have better access to their internal state, and because we can take steps to certify that the machine's explanations are not deceptions (intentional or self-deception), something we can never do for sure 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 we saw 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 it—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. To gain trust that they are not biased would require not an explanation, but rather an **audit** of past decisions, with aggregated statistics across various demographic groups, to see if their approval rates are balanced.

27.3.5 The future of work

From the first agricultural revolution (10,000 BCE) to the industrial revolution (1760s) to the green revolution in food production (1950s), new technologies have changed the way humanity works and lives. AI technology has played an increasing role in the global economy, reaching $1 billion in annual revenue in the 1980s and surpassing $1 trillion in 2018, according to Gartner Inc, who also predict $4 trillion by 2022. Opinions vary: PwC (Rao and Verweij, 2017) predicts $15 trillion by 2030, and McKinsey suggests a near-term target of $6 trillion, without specifying an exact year. The gains in wealth come from a combination of automation, increased productivity for existing workers, and increased consumer demand for new products. 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 expected.

Technological innovations have often put some people out of work: weavers were replaced by automated looms in the 1810s (leading to the Luddite protests); farm laborers were replaced by machinery starting in the 1930s (leading John Maynard Keynes to coin the term **technological unemployment**); bank tellers were replaced by ATMs in the 2000s. In the past, increased productivity has always led to increased wealth and increased demand, and thus net job growth. Each bank branch required fewer tellers, but that made it cheaper to operate a branch, and the number of branches increased. This increased the number of bank employees, but most of the new ones were performing more advanced tasks, such as arranging loans, rather than counting out change. We see a consistant pattern that automation has led to increased productivity by making certain *tasks* easier, without causing a large change in the number of *jobs*.

Technological unemployment

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 automation. But some 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. Brynjolfsson and McAfee (2014) point out that the first industrial revolution introduced machines that were *complements* to human labor, serving to amplify human productivity, but that the machines of the AI age are often *substitutes* for human work. 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 at risk of being replaced by driverless cars/trucks/buses/taxis.

Tasks

It is important to make a distinction 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, in the future a truck driver will spend less time holding the steering wheel, and more time making sure that the goods are picked up and delivered properly; serving as a customer service representative and salesperson at either end of the journey; and perhaps serving as the manager of a convoy of, say, three robotic trucks. Replacing three drivers with one 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 not all of them.

Business process automation

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.

As we face an aging population in almost all developed countries, the mix between workers and retirees changes—in 2015 we have less than 30 people over the age of 65 for every 100 workers; projections are that by 2050 there will be over 60 per 100 workers. Care for the elderly will be an increasingly important role, one that can partially be filled by AI, and 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.

Pace of change

Even if automation has a multi-trillion-dollar net positive impact, there may still be problems due to the **pace of change**. LinkedIn founder Reid Hoffman commented that "Transitions can be very painful. Let's try to make it work out in a way that's more humane." 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 is a huge disruption in

[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.

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.

Income inequality

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. This winner-take-all aspect is clearly bad for the losers, but it is stressful for the winners as well, who know that if they slack off even momentarily, they may be passed by a competitor. 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.

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 and accomplishment. With increasing automation, it may be that these three purposes become disaggregated—society's needs will be served by automation, and in the long run, individuals get their sense of purpose from contributions outside of a job. Their needs can be fulfilled by social policies that include a combination of free or inexpensive access to social services and education, portable health care, retirement, and education accounts, trade adjustment assistance, progressive tax rates, earned income tax credits, negative income tax, or universal basic income.

27.3.6 Robot rights

The question of robot consciousness, discussed in Section 27.2, 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 "persons," then the argument can be made (e.g., by Sparrow (2004)) that they have human 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 *A.I.* 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 the real world, 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. Joanna Bryson, a leading anti-robot-rights advocate, argues (Bryson, 2010) that calling robots "human" would dehumanize actual people. 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.

27.3.7 AI Safety

Almost any technology has the potential to cause harm in the wrong hands, but with AI and robotics, the hands might belong to the technology itself. Countless science fiction stories have warned about robots or cyborgs running amok. Early examples include Mary Shelley's *Frankenstein, or the Modern Prometheus* (1818)[4] 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 release 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

Failure modes and effect analysis (FMEC)

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 (FMEC)**: 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

[4] As a young man, Charles Babbage was influenced by reading *Frankenstein*.

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 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.

Fault tree analysis (FTA)

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. 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. However, 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 off explicit programming, machine learning over time, and adherence to low impact.

Unintended side effects

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):

Externalities

Tragedy of the commons

- 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. Other agents took advantage of a pattern in the file names of training examples to generate a solution that performs flawlessly on training data and poorly on real data. A genetic algorithm tasked to design a sorting function came up with a simple function that always returned the empty list, regardless of the input. This worked because the designers specified the goal as "the output list is sorted" and forgot to specify that "the output is a permutation of the input." 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. Designers of agents should make sure their systems do not fall victim to the attacks on this list; 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 33. We run into trouble whan 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 kidnapping said person is not an option. But a robotic cleaner may not know this unless we remember to tell it. Or, consider Nick Bostrom's (2014) scenario where a robot factory designed to create paper clips ends up capturing all of Earth's natural resources in order to maximize its output. The theory of this scenario has been discussed in depth, but in practice it has not yet been a problem. That is not only because agents are not yet clever enough to carry out such diabolical plans, but also because we train machine learning systems from examples, and the training data never contains plans like that—in effect, the training data embodies the societal norms.

In addition to striving to get the utility function right, we should also admit that we may get it wrong, and build an agent that can deal with the inevitable errors. One issue is that desires change over time. For example, if technology had allowed us to design an AI agent in 1800 and endow it with the prevailing morals of the time, it might be fighting today to reestablish slavery and abolish women's right to vote. This suggests that we allow the utility function to evolve over time. This could be done by human designers periodically updating the software, but then we have to rely on the update schedule, and we have to carefully test and verify each update.

An alternative is to let the agent update its own utility function, but that in itself can be risky. We don't want the agent to reason: "Humans think it is moral to kill annoying insects, in part because insect brains are so primitive. But human brains are primitive compared to my powers, so it must be moral for me to kill humans." We can design an agent to learn by observing examples of human behavior—what is called apprenticeship learning

(Section 22.3.3). Part of the apprenticeship is learning what actions to take in what situations, and part is learning the utility function that the humans must be operating under, using **inverse reinforcement learning**. This approach can exceed human-level performance by smoothing out actions and eliminating errors, as in the work by Coates *et al.* (2009) on learning to fly autonomous helicopters by observing the actions of human pilots. If we program our robots with a strong prior to follow past behavior examples then it is unlikely that they will start doing something wildly inappropriate. However, we also lessen the possibility that they start doing something wildly great. For example, ALPHAGO was trained on human Go games and learned to play very appropriately, beating world champions. But ALPHAZERO was trained only through self-play with no examples of human behaviors, and it eventually learned to play even better—it was not constrained by past human performance.

We could also allow a robot to ask a human when it is uncertain: "would you like me to convert all of Earth's matter into paper clips?" But humans may not always be able to answer such questions wisely—they may not be able to foresee all the ramifications of an action, they 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. Humans have some innately aggressive tendencies. Part of AI safety is to avoid these problems. The machines we build need not be innately aggressive, unless we decide to build them that way (or unless they emerge as the end product of a mechanism design that encourages aggressive behavior). The idea of safeguards against improper desires is not a new one. In the *Odyssey*, (ca. 700 BCE) described ' encounter with the sirens, whose song was so alluring it compelled sailors to cast themselves into the sea. Knowing the song would have that effect on him, but longing to hear it, Ulysses ordered his crew to bind him to the mast so that he could not perform the self-destructive act. It is interesting to think how similar safeguards could be built into AI systems. The challenge is one of mechanism design: what are the mechanisms for allowing change to the utility function, and how will such changes play out, even in the face of adversaries?

Despite this toolbox of safeguards, there is a fear, expressed by prominent technologists such as Bill Gates, Stephen Hawking, Elon Musk 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. We have seen that when there is a balance of power—no one entity has a huge advantage over all the others—then the worst behavior can be checked. But the history of wars and corporate malfeasance should give us pause. Moreover, AI systems may be different because of their potential to rapidly self-improve, 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, 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 (consider Moore's Law). However, it is a leap to extrapolate all the way 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. The history of flight progressed dramatically from the Wright brothers' flight in 1903 to the moon landing in 1969, but has had no breakthroughs of similar magnitude since then.

We do know there are limits on computability and computational complexity. If the problem of defining ultraintelligent machines happens to fall in the class of, say, NEXPTIME-complete problems, and if there are no heuristic shortcuts, then even exponential progress in technology won't help—the speed of light puts a strict upper bound on how much computing can be done, and problems beyond that limit will not be solved.

Thinkism

Wicked problems

Furthermore, the whole idea of an ultraintelligent machine leading to a technological singularity assumes that intelligence is an especially important attribute, and if you have enough of it, all problems can be solved. Kevin Kelly calls this assumption **thinkism**—a prejudice in favor of *thinking* over all other abilities. It is true that many problems, such as playing Go, can be conquered by intelligent algorithms. But so-called **wicked problems** (Rittel and Webber, 1973)—with underspecified, contradictory, ever-changing requirements and missing data—may not fall so easily.

An ultraintelligent machine tasked with creating a theory of dark matter 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 Go, and how many are like the Middle East.

Transhumanism

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):

> The Singularity will allow us to transcend these limitations of our biological bodies and brain. We will gain power over our fates. Our mortality will be in our own hands. We will be able to live as long as we want (a subtly different statement from saying we will live forever). 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 and trustworthy.
- 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 2016–2019. The **Chinese Room** has been debated endlessly (Searle, 1980; Chalmers, 1992; Preston and Bishop, 2002).

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).[5] 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 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 Weiner'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. Weiner 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 considered issues of AI ethics 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 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

[5] Science fiction writers are in broad agreement that robots are very bad at resolving contradictions. In *2001*, the HAL 9000 robot 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.

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. Armin Krishnan (2016) and Paul Scharre (2018) offer more recent books on the same topic. 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 Dwork *et al.* (2014) give practical examples of clever ways to apply differential privacy to get better results than the naive approach. 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) is a foundational paper that shows 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.* (2016) 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) advocate for fair impact: all classes should have the same expected utility. Liu *et al.* (2018) 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.

Trust: Explainable AI was an important topic going back to the days of expert systems (Neches *et al.*, 1985), and is 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. 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* (2015) 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 personify 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, which then falls apart when the machine inevitably fails. Norbert Weiner'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.