26.3 The Ethics and Risks of Developing Artificial Intelligence

So far, we have concentrated on whether we *can* develop AI, but we must also consider whether we *should*. If the effects of AI technology are more likely to be negative than positive, then it would be the moral responsibility of workers in the field to redirect their research. Many new technologies have had unintended negative side effects: nuclear fission brought Chernobyl and the threat of global destruction; the internal combustion engine brought air pollution, global warming, and the paving-over of paradise. In a sense, automobiles are robots that have conquered the world by making themselves indispensable.

All scientists and engineers face ethical considerations of how they should act on the job, what projects should or should not be done, and how they should be handled. See the handbook on the *Ethics of Computing* (Berleur and Brunnstein, 2001). AI, however, seems to pose some fresh problems beyond that of, say, building bridges that don't fall down:

- People might lose their jobs to automation.
- People might have too much (or too little) leisure time.
- People might lose their sense of being unique.
- AI systems might be used toward undesirable ends.
- The use of AI systems might result in a loss of accountability.
- The success of AI might mean the end of the human race.

We will look at each issue in turn.

People might lose their jobs to automation. The modern industrial economy has become dependent on computers in general, and select AI programs in particular. For example, much of the economy, especially in the United States, depends on the availability of consumer credit. Credit card applications, charge approvals, and fraud detection are now done by AI programs. One could say that thousands of workers have been displaced by these AI programs, but in fact if you took away the AI programs these jobs would not exist, because human labor would add an unacceptable cost to the transactions. So far, automation through information technology in general and AI in particular has created more jobs than it has eliminated, and has created more interesting, higher-paying jobs. Now that the canonical AI program is an "intelligent agent" designed to assist a human, loss of jobs is less of a concern than it was when AI focused on "expert systems" designed to replace humans. But some researchers think that doing the complete job is the right goal for AI. In reflecting on the 25th Anniversary of the AAAI, Nils Nilsson (2005) set as a challenge the creation of human-level AI that could pass the employment test rather than the Turing Test—a robot that could learn to do any one of a range of jobs. We may end up in a future where unemployment is high, but even the unemployed serve as managers of their own cadre of robot workers.

People might have too much (or too little) leisure time. Alvin Toffler wrote in *Future Shock* (1970), "The work week has been cut by 50 percent since the turn of the century. It is not out of the way to predict that it will be slashed in half again by 2000." Arthur C. Clarke (1968b) wrote that people in 2001 might be "faced with a future of utter boredom, where the main problem in life is deciding which of several hundred TV channels to select."

The only one of these predictions that has come close to panning out is the number of TV channels. Instead, people working in knowledge-intensive industries have found themselves part of an integrated computerized system that operates 24 hours a day; to keep up, they have been forced to work *longer* hours. In an industrial economy, rewards are roughly proportional to the time invested; working 10% more would tend to mean a 10% increase in income. In an information economy marked by high-bandwidth communication and easy replication of intellectual property (what Frank and Cook (1996) call the "Winner-Take-All Society"), there is a large reward for being slightly better than the competition; working 10% more could mean a 100% increase in income. So there is increasing pressure on everyone to work harder. 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.

People might lose their sense of being unique. In *Computer Power and Human Reason*, Weizenbaum (1976), the author of the ELIZA program, points out some of the potential threats that AI poses to society. One of Weizenbaum's principal arguments is that AI research makes possible the idea that humans are automata—an idea that results in a loss of autonomy or even of humanity. We note that the idea has been around much longer than AI, going back at least to *L'Homme Machine* (La Mettrie, 1748). Humanity has survived other setbacks to our sense of uniqueness: *De Revolutionibus Orbium Coelestium* (Copernicus, 1543) moved the Earth away from the center of the solar system, and *Descent of Man* (Darwin, 1871) put *Homo sapiens* at the same level as other species. AI, if widely successful, may be at least as threatening to the moral assumptions of 21st-century society as Darwin's theory of evolution was to those of the 19th century.

AI systems might be used toward undesirable ends. Advanced technologies have often been used by the powerful to suppress their rivals. As the number theorist G. H. Hardy wrote (Hardy, 1940), "A science is said to be useful if its development tends to accentuate the existing inequalities in the distribution of wealth, or more directly promotes the destruction of human life." This holds for all sciences, AI being no exception. Autonomous 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 Iraq (Singer, 2009). One moral theory holds that military robots are like medieval armor taken to its logical extreme: no one would have moral objections to a soldier wanting to wear a helmet when being attacked by large, angry, axe-wielding enemies, and a teleoperated robot is like a very safe form of armor. On the other hand, robotic weapons pose additional risks. To the extent that human decision making is taken out of the firing loop, robots may end up making decisions that lead to the killing of innocent civilians. At a larger scale, the possession of powerful robots (like the possession of sturdy helmets) may give a nation overconfidence, causing it to go to war more recklessly than necessary. In most wars, at least one party is overconfident in its military abilities—otherwise the conflict would have been resolved peacefully.

Weizenbaum (1976) also pointed out that speech recognition technology could lead to widespread wiretapping, and hence to a loss of civil liberties. He didn't foresee a world with terrorist threats that would change the balance of how much surveillance people are willing to

accept, but he did correctly recognize that AI has the potential to mass-produce surveillance. His prediction has in part come true: the U.K. now has an extensive network of surveillance cameras, and other countries routinely monitor Web traffic and telephone calls. Some accept that computerization leads to a loss of privacy—Sun Microsystems CEO Scott McNealy has said "You have zero privacy anyway. Get over it." David Brin (1998) argues that loss of privacy is inevitable, and the way to combat the asymmetry of power of the state over the individual is to make the surveillance accessible to all citizens. Etzioni (2004) argues for a balancing of privacy and security; individual rights and community.

The use of AI systems might result in a loss of accountability. In the litigious atmosphere that prevails in the United States, legal liability becomes an important issue. When a physician relies on the judgment of a medical expert system for a diagnosis, who is at fault if the diagnosis is wrong? Fortunately, due in part to the growing influence of decision-theoretic methods in medicine, it is now accepted that negligence cannot be shown if the physician performs medical procedures that have high *expected* utility, even if the *actual* result is catastrophic for the patient. The question should therefore be "Who is at fault if the diagnosis is unreasonable?" So far, courts have held that medical expert systems play the same role as medical textbooks and reference books; physicians are responsible for understanding the reasoning behind any decision and for using their own judgment in deciding whether to accept the system's recommendations. In designing medical expert systems as agents, therefore, the actions should be thought of not as directly affecting the patient but as influencing the physician's behavior. If expert systems become reliably more accurate than human diagnosticians, doctors might become legally liable if they *don't* use the recommendations of an expert system. Atul Gawande (2002) explores this premise.

Similar issues are beginning to arise regarding the use of intelligent agents on the Internet. Some progress has been made in incorporating constraints into intelligent agents so that they cannot, for example, damage the files of other users (Weld and Etzioni, 1994). The problem is magnified when money changes hands. If monetary transactions are made "on one's behalf" by an intelligent agent, is one liable for the debts incurred? Would it be possible for an intelligent agent to have assets itself and to perform electronic trades on its own behalf? So far, these questions do not seem to be well understood. To our knowledge, no program has been granted legal status as an individual for the purposes of financial transactions; at present, it seems unreasonable to do so. Programs are also not considered to be "drivers" for the purposes of enforcing traffic regulations on real highways. In California law, at least, there do not seem to be any legal sanctions to prevent an automated vehicle from exceeding the speed limits, although the designer of the vehicle's control mechanism would be liable in the case of an accident. As with human reproductive technology, the law has yet to catch up with the new developments.

The success of AI might mean the end of the human race. Almost any technology has the potential to cause harm in the wrong hands, but with AI and robotics, we have the new problem that the wrong hands might belong to the technology itself. Countless science fiction stories have warned about robots or robot–human cyborgs running amok. Early examples

include Mary Shelley's *Frankenstein, or the Modern Prometheus* (1818)[5] and Karel Capek's play *R.U.R.* (1921), in which robots conquer the world. In movies, we have *The Terminator* (1984), which combines the cliches of robots-conquer-the-world with time travel, and *The Matrix* (1999), which combines robots-conquer-the-world with brain-in-a-vat.

It seems that robots are the protagonists of so many conquer-the-world stories because they represent the unknown, just like the witches and ghosts of tales from earlier eras, or the Martians from *The War of the Worlds* (Wells, 1898). The question is whether an AI system poses a bigger risk than traditional software. We will look at three sources of risk.

First, the AI system's state estimation may be incorrect, causing it to do the wrong thing. For example, an autonomous car might incorrectly estimate the position of a car in the adjacent lane, leading to an accident that might kill the occupants. More seriously, a missile defense system might erroneously detect an attack and launch a counterattack, leading to the death of billions. These risks are not really risks of AI systems—in both cases the same mistake could just as easily be made by a human as by a computer. The correct way to mitigate these risks is to design a system with checks and balances so that a single state-estimation error does not propagate through the system unchecked.

Second, specifying the right utility function for an AI system to maximize is not so easy. For example, we might propose a utility function designed to *minimize human suffering*, expressed as an additive reward function over time as in Chapter 17. Given the way humans are, however, we'll always find a way to suffer even in paradise; so the optimal decision for the AI system is to terminate the human race as soon as possible—no humans, no suffering. With AI systems, then, we need to be very careful what we ask for, whereas humans would have no trouble realizing that the proposed utility function cannot be taken literally. On the other hand, computers need not be tainted by the irrational behaviors described in Chapter 16. Humans sometimes use their intelligence in aggressive ways because humans have some innately aggressive tendencies, due to natural selection. 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). Fortunately, there are techniques, such as apprenticeship learning, that allows us to specify a utility function by example. One 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.

Third, the AI system's learning function may cause it to evolve into a system with unintended behavior. This scenario is the most serious, and is unique to AI systems, so we will cover it in more depth. I. J. Good wrote (1965),

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.

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

TECHNOLOGICAL SINGULARITY

The "intelligence explosion" has also been called the **technological singularity** by mathematics professor and science fiction author Vernor Vinge, who writes (1993), "Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended." Good and Vinge (and many others) correctly note that the curve of technological progress (on many measures) is growing exponentially at present (consider Moore's Law). However, it is a leap to extrapolate that the curve will continue to a singularity of near-infinite growth. So far, every other technology has followed an S-shaped curve, where the exponential growth eventually tapers off. Sometimes new technologies step in when the old ones plateau; sometimes we hit hard limits. With less than a century of high-technology history to go on, it is difficult to extrapolate hundreds of years ahead.

Note that the concept of ultraintelligent machines assumes that intelligence is an especially important attribute, and if you have enough of it, all problems can be solved. But we know there are limits on computability and computational complexity. If the problem of defining ultraintelligent machines (or even approximations to them) 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; problems beyond that limit will not be solved. We still don't know where those upper bounds are.

Vinge is concerned about the coming singularity, but some computer scientists and futurists relish it. Hans Moravec (2000) encourages us to give every advantage to our "mind children," the robots we create, which may surpass us in intelligence. There is even a new word—**transhumanism**—for the active social movement that looks forward to this future in which humans are merged with—or replaced by—robotic and biotech inventions. Suffice it to say that such issues present a challenge for most moral theorists, who take the preservation of human life and the human species to be a good thing. Ray Kurzweil is currently the most visible advocate for the singularity view, writing in *The Singularity is Near* (2005):

TRANSHUMANISM

> The Singularity will allow us to transcend these limitations of our biological bodies and brain. We will gain power over our fates. 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.

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."

If ultraintelligent machines are a possibility, we humans would do well to make sure that we design their predecessors in such a way that they design themselves to treat us well. Science fiction writer Isaac Asimov (1942) was the first to address this issue, with his three laws of robotics:

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.

These laws seem reasonable, at least to us humans.[6] But the trick is how to implement these laws. In the Asimov story *Roundabout* a robot is sent to fetch some selenium. Later the robot is found wandering in a circle around the selenium source. Every time it heads toward the source, it senses a danger, and the third law causes it to veer away. But every time it veers away, the danger recedes, and the power of the second law takes over, causing it to veer back towards the selenium. The set of points that define the balancing point between the two laws defines a circle. This suggests that the laws are not logical absolutes, but rather are weighed against each other, with a higher weighting for the earlier laws. Asimov was probably thinking of an architecture based on control theory—perhaps a linear combination of factors—while today the most likely architecture would be a probabilistic reasoning agent that reasons over probability distributions of outcomes, and maximizes utility as defined by the three laws. But presumably we don't want our robots to prevent a human from crossing the street because of the nonzero chance of harm. That means that the negative utility for harm to a human must be much greater than for disobeying, but that each of the utilities is finite, not infinite.

FRIENDLY AI

Yudkowsky (2008) goes into more detail about how to design a **Friendly AI**. He asserts that friendliness (a desire not to harm humans) should be designed in from the start, but that the designers should recognize both that their own designs may be flawed, and that the robot will learn and evolve over time. Thus the challenge is one of mechanism design—to define a mechanism for evolving AI systems under a system of checks and balances, and to give the systems utility functions that will remain friendly in the face of such changes.

We can't just give a program a static utility function, because circumstances, and our desired responses to circumstances, change over time. For example, if technology had allowed us to design a super-powerful AI agent in 1800 and endow it with the prevailing morals of the time, it would be fighting today to reestablish slavery and abolish women's right to vote. On the other hand, if we build an AI agent today and tell it to evolve its utility function, how can we assure that it won't reason that "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."

Omohundro (2008) hypothesizes that even an innocuous chess program could pose a risk to society. Similarly, Marvin Minsky 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 to help achieve its goal. The moral is that even if you only want your program to play chess or prove theorems, if you give it the capability to learn and alter itself, you need safeguards. Omohundro concludes that "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," This seems to be an excellent idea for society in general, regardless of the possibility of ultraintelligent machines.

[6] A robot might notice the inequity that a human is allowed to kill another in self-defense, but a robot is required to sacrifice its own life to save a human.

We should note that the idea of safeguards against change in utility function is not a new one. In the *Odyssey*, Homer (ca. 700 B.C.) described Ulysses' encounter with the sirens, whose song was so alluring it compelled sailors to cast themselves into the sea. Knowing it would have that effect on him, 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.

Finally, let us consider the robot's point of view. If robots become conscious, then to treat them as mere "machines" (e.g., to take them apart) might be immoral. Science fiction writers have addressed the issue of robot rights. The movie *A.I.* (Spielberg, 2001) was based on a story by Brian Aldiss about an intelligent robot who was programmed to believe that he was human and fails to understand his eventual abandonment by his owner–mother. The story (and the movie) argue for the need for a civil rights movement for robots.

26.4 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.
- There is general agreement in modern times that mental states are brain states.
- Arguments for and against strong AI are inconclusive. Few mainstream AI researchers believe that anything significant hinges on the outcome of the debate.
- Consciousness remains a mystery.
- We identified six potential threats to society posed by AI and related technology. We concluded that some of the threats are either unlikely or differ little from threats posed by "unintelligent" technologies. One threat in particular is worthy of further consideration: that ultraintelligent machines might lead to a future that is very different from today—we may not like it, and at that point we may not have a choice. Such considerations lead inevitably to the conclusion that we must weigh carefully, and soon, the possible consequences of AI research.

BIBLIOGRAPHICAL AND HISTORICAL NOTES

Sources for the various responses to Turing's 1950 paper and for the main critics of weak AI were given in the chapter. Although it became fashionable in the post-neural-network era