EXERCISES 28

PHILOSOPHY, ETHICS, AND SAFETY OF AI

28.1 The Limits of AI

Exercise 28.DISA

Go through Turing's list of alleged "disabilities" of machines, identifying which have been achieved, which are achievable in principle by a program, and which are still problematic because they require conscious mental states.

We will take the disabilities (see page 1000) one at a time. Note that this exercise might be better as a class discussion rather than written work.

a. *be kind*: Certainly there are programs that are polite and helpful, but to be kind requires an intentional state, so this one is problematic.

b. *resourceful*: Resourceful means "clever at finding ways of doing things." Many programs meet this criteria to some degree: a compiler can be clever making an optimization that the programmer might not ever have thought of; a database program might cleverly create an index to make retrievals faster; a checkers or backgammon program learns to play as well as any human. One could argue whether the machines are "really" clever or just seem to be, but most people would agree this requirement has been achieved.

c. *beautiful*: Its not clear if Turing meant to be beautiful or to create beauty, nor is it clear whether he meant physical or inner beauty. Certainly the many industrial artifacts in the New York Museum of Modern Art, for example, are evidence that a machine can be beautiful. There are also programs that have created art. One of the best known of these is chronicled in *Aaron's code: Meta-art, artificial intelligence, and the work of Harold Cohen* (McCorduck, 1991). There have been many instances of computers making art since, such as with the music of David Cope (Cope, 2008) or the images of Google's Deep Dream (Mordvinstev, 2015).

d. *friendly* This appears to fall under the same category as *kind*.

e. *have initiative* Interestingly, there is now a serious debate whether software should take initiative. The whole field of software agents says that it should; critics such as Ben Schneiderman say that to achieve predictability, software should only be an assistant, not an autonomous agent. Notice that the debate over whether software *should* have initiative presupposes that it *has* initiative.

f. *have a sense of humor* We know of no major effort to produce humorous works. However, this seems to be achievable in principle. All it would take is someone like Harold Cohen who is willing to spend a long time tuning a humor-producing machine. We note that humorous text is probably easier to produce than other media. For a recent treatment of humor, its mechanisms, and biological relevancy, see Hurley, 2011.

g. *tell right from wrong* There is considerable research in applying AI to legal reasoning, and there are now tools that assist the lawyer in deciding a case and doing research. One could argue whether following legal precedents is the same as telling right from wrong, and in any case this has a problematic conscious aspect to it.

h. *make mistakes* At this stage, every computer user is familiar with software that makes mistakes! It is interesting to think back to what the world was like in Turing's day, when some people thought it would be difficult or impossible for a machine to make mistakes.

i. *fall in love* This is one of the cases that clearly requires consciousness. Note that while some people claim that their pets love them, and some claim that pets are not conscious, we don't know of anybody who makes both claims.

j. *enjoy strawberries and cream* There are two parts to this. First, there has been little to no work on taste perception in AI , so we're nowhere near a breakthrough on this. Second, the "enjoy" part clearly requires consciousness.

k. *make someone fall in love with it* This criteria is actually not too hard to achieve; machines such as dolls and teddy bears have been doing it to children for centuries. Machines that talk and have more sophisticated behaviors just have a larger advantage in achieving this.

l. *learn from experience* Part VI shows that this has been achieved many times in AI.

m. *use words properly* No program uses words perfectly, but there have been many natural language programs that use words properly and effectively within a limited domain (see Chapters 23-24).

n. *be the subject of its own thought* The problematic word here is "thought." Many programs can process themselves, as when a compiler compiles itself. Perhaps closer to human self-examination is the case where a program has an imperfect representation of itself. One anecdote of this involves Doug Lenat's Eurisko program. It used to run for long periods of time, and periodically needed to gather information from outside sources. It "knew" that if a person were available, it could type out a question at the console, and wait for a reply. Late one night it saw that no person was logged on, so it couldn't ask the question it needed to know. But it knew that Eurisko itself was up and running, and decided it would modify the representation of Eurisko so that it inherits from "Person," and then proceeded to ask itself the question!

o. *have as much diversity of behavior as man* Clearly, no machine has achieved this, although there is no principled reason why one could not.

p. *do something really new* This seems to be just an extension of the idea of learning from experience: if you learn enough, you can do something really new. "Really" is subjective, and some would say that no machine has achieved this yet. On the other hand, professional Go players believe that AlphaZero has revolutionized the theory of the game of Go.

Exercise 28.POPM

Find and analyze an account in the popular media of one or more of the arguments to the effect that AI is impossible.

This exercise depends on what happens to have been published lately. Here is an excerpt from a reviews of Roger Penrose's book *Shadows of the Mind*, by Adam Schulman (1995):

> Roger Penrose, the distinguished mathematical physicist, has again entered the lists to rid the world of a terrible dragon. The name of this dragon is "strong artificial intelligence."
>
> Strong Al, as its defenders call it, is both a widely held scientific thesis and an ongoing technological program. The thesis holds that the human mind is nothing but a fancy calculating machine-"-a computer made of meat"–and that all thinking is merely computation; the program is to build faster and more powerful computers that will eventually be able to do everything the human mind can do and more. Penrose believes that the thesis is false and the program unrealizable, and he is confident that he can prove these assertions. . . .
>
> In Part I of Shadows of the Mind Penrose makes his rigorous case that human consciousness cannot be fully understood in computational terms. . . . How does Penrose prove that there is more to consciousness than mere computation? Most people will already find it inherently implausible that the diverse faculties of human consciousness–self-awareness, understanding, willing, imagining, feeling–differ only in complexity from the workings of, say, an IBM PC.

Students should have no problem finding things in this and other articles with which to disagree.

Dubious claims also emerge from the interaction between journalists' desire to write entertaining and controversial articles and academics' desire to achieve prominence and to be viewed as ahead of the curve. Here's one typical result— *Is Nature's Way The Best Way?*, Omni, February 1995, p. 62:

> Artificial intelligence has been one of the least successful research areas in computer science. That's because in the past, researchers tried to apply conventional computer programming to abstract human problems, such as recognizing shapes or speaking in sentences. But researchers at MIT's Media Lab and Boston University's Center for Adaptive Systems focus on applying paradigms of intelligence closer to what nature designed for humans, which include evolution, feedback, and adaptation, are used to produce computer programs that communicate among themselves and in turn learn from their mistakes. *Profiles In Artificial Intelligence, David Freedman.*

This is not an argument that AI is impossible, just that it has been unsuccessful. The full text of the article is not given, but it is implied that the argument is that evolution worked for humans, therefore it is a better approach for programs than is "conventional computer programming." This is a common argument, but one that ignores the fact that (a) there are many possible solutions to a problem; one that has worked in the past may not be the best in the present (b) we don't have a good theory of evolution, so we may not be able to duplicate human evolution, (c) natural evolution takes millions of years and for almost all animals does not result in intelligence; there is no guarantee that artificial evolution will do better (d) artificial evolution (or genetic algorithms, ALife, neural nets, etc.) is not the only approach that involves feedback, adaptation and learning. "Conventional" AI does this as well.

28.2 Can Machines Really Think?

Exercise 28.DEFI

Attempt to write definitions of the terms "intelligence," "thinking," and "consciousness." Suggest some possible objections to your definitions.

This also might make a good class discussion topic. A consultation of the literature will divulge many recent attempts, such as that of Chollet, 2019 on intelligence. Here are our attempts:

intelligence: a measure of the ability of an agent to make the right decisions, given the available evidence. Given the same circumstances, a more intelligent agent will make better decisions on average.

thinking: creating internal representations in service of the goal of coming to a conclusion, making a decision, or weighing evidence.

consciousness: being aware of one's own existence, and of one's current internal state.

Here are some objections (with replies):

For **intelligence**, too much emphasis is put on decision-making. Haven't you ever known a highly intelligent person who made bad decisions? Also no mention is made of learning. You can't be intelligent by using brute-force look-up, for example, could you? [The emphasis on decision-making is only a liability when you are working at too coarse a granularity (e.g., "What should I do with my life?") Once you look at smaller-grain decisions (e.g., "Should I answer a, b, c or none of the above?"), you get at the kinds of things tested by current IQ tests, while maintaining the advantages of the action-oriented approach covered in Chapter 1. As to the brute-force problem, think of intelligence in terms of an ecological niche: an agent only needs to be as intelligent as is necessary to be successful. If this can be accomplished through some simple mechanism, fine. For the complex environments that we humans are faced with, more complex mechanisms are needed.]

For **thinking**, we have the same objections about decision-making, but in general, thinking is the least controversial of the three terms.

For **consciousness**, the weakness is the definition of "aware." How does one demonstrate awareness? Also, it is not one's true internal state that is important, but some kind of abstraction or representation of some of the features of it. Deacon, 2011 offers a novel physical argument for biological naturalism with implications that machines cannot have consciousness. Those particularly interested in this subject might consider attempting to support or refute his argument using the formalism of AI.

Exercise 28.CHRM

Does a refutation of the Chinese room argument necessarily prove that appropriately programmed computers have mental states? Does an acceptance of the argument necessarily mean that computers cannot have mental states?

No. Searle's Chinese room thesis says that there are some cases where running a program that generates the right output for the Chinese room does not cause true understanding/consciousness. The negation of this thesis is therefore that all programs with the right output do cause true understanding/consciousness. So if you were to disprove Searle's *thesis*, then you would have a proof of machine consciousness. However, what this question is getting at is the *argument* behind the thesis. If you show that the argument is faulty, then you may have proved nothing more: it might be that the thesis is true (by some other argument), or it might be false.

Exercise 28.GODI

Alan Perlis (1982) wrote, "A year spent in artificial intelligence is enough to make one believe in God". He also wrote, in a letter to Philip Davis, that one of the central dreams of computer science is that "through the performance of computers and their programs we will remove all doubt that there is only a chemical distinction between the living and nonliving world." To what extent does the progress made so far in artificial intelligence shed light on these issues? Suppose that at some future date, the AI endeavor has been completely successful; that is, we have build intelligent agents capable of carrying out any human cognitive task at human levels of ability. To what extent would that shed light on these issues?

The progress that has been made so far — a limited class of restricted cognitive activities can be carried out on a computer, some much better than humans, most much worse than humans — is very little evidence. If all cognitive activities can be explained in computational terms, then that would at least establish that cognition does not *require* the involvement of anything beyond physical processes. Of course, it would still be possible that something of the kind is *actually* involved in human cognition, but this would certainly increase the burden of proof on those who claim that it is.

Exercise 28.GOOD

I. J. Good claims that intelligence is the most important quality, and that building ultraintelligent machines will change everything. A sentient cheetah counters that "Actually speed is more important; if we could build ultrafast machines, that would change everything," and a sentient elephant claims "You're both wrong; what we need is ultrastrong machines." What do you think of these arguments?

This question asks whether our obsession with intelligence merely reflects our view of ourselves as distinct due to our intelligence. One may respond in two ways. First, note that we already have ultrafast and ultrastrong machines (for example, aircraft and cranes) but they have not changed everything—only those aspects of life for which raw speed and strength are important. Good's argument is based on the view that intelligence is important in all aspects of life, since all aspects involve choosing how to act. Second, note that ultraintelligent machines have the special property that they can easily create ultrafast and ultrastrong machines as needed, whereas the converse is not true.

Exercise 28.IMPO

Some critics object that AI is impossible, while others object that it is *too* possible and that ultraintelligent machines pose a threat. Which of these objections do you think is more likely? Would it be a contradiction for someone to hold both positions?

To decide if AI is impossible, we must first define it. In this book, we've chosen a definition that makes it easy to show it is possible in theory—for a given architecture, we just enumerate all programs and choose the best. In practice, this might still be infeasible, but recent history shows steady progress at a wide variety of tasks. Now if we define AI as the production of agents that act indistinguishably form (or at least as intellgiently as) human beings on any task, then one would have to say that little progress has been made, and some, such as Marvin Minsky, bemoan the fact that few attempts are even being made. Others think it is quite appropriate to address component tasks rather than the "whole agent" problem. Our feeling is that AI is neither impossible nor a looming threat. But it would be perfectly consistent for someone to feel that AI is most likely doomed to failure, but still that the risks of possible success are so great that it should not be pursued for fear of success.

28.3 The Ethics of AI

Exercise 28.WEAP

Why might state actors support or oppose autonomous weapons, as discussed in Section **??**? What role can technologists play in this debate?

Arguments **in support** might include: a desire to adhere to international law, such as the Martens Clause to the Geneva convention which bans bans weapons that violate the "principles of humanity and the dictates of public conscience" (Strebel, 1982); a desire to establish treaties regulating arms in lieu of an arms race and the security threat such a race might impose, as has occurred with nuclear weapons and with biological agents under the Nixon administration in the United States (as, for example, argued in Russell, 2015 and discussed in Section **??**). Of course there will be some parties who do not adhere to such treaties (such as non state entities), but these will likely be of negligible impact compared to that of large nations, although recall the **dual use** problem described in the text.

This presages arguments **in opposition**: dictators, terrorists groups, and repressive regimes may want access to cheaper and deadlier weapons for obvious reasons; even established nations may want to reduce their own casualties (and training costs) by reducing battlefield exposure of soldiers; and autonomous weapons may even be able to reduce casualties overall under the assumption that because the threat to an autonomous weapon is small compared to the threat to a human (presumably we value a lost robot much less than a lost life), the weapon can wait to see how a dangerous situation plays out in a risky or otherwise unknown situation. Robots don't have to shoot first.

As evident in previous arms control movements, technologists, or those with an intimate knowledge of the weapon technologies, can play quite significant roles. For example, physi-

cist Robert Oppenheimer, who was instrumental in the production of the first atomic bomb, famously became an outspoken opponent of nuclear proliferation after the conclusion of the Second World War. Computer scientists have done the same, such as Norbert Wiener, one of the founders of control theory and famous for his work on anti-aircraft guns, who spoke out against the militarization of computing technology after the conclusion of WWII. The now-defunct Computing Professionals for Social Responsibility and many other organizations have continued in Wiener's legacy. This legacy is notably similar to the expert recommendations with regard to the appropriate use of technology as advocated by the authors of this book. Similar trends have cropped up in the late 2010s movement at major computing companies in which employees began to debate the utility of furnishing such military contracts for government agencies (such as the U.S. DoD). The effects of these actions are not yet apparent and will be borne out in time.

Exercise 28.NANO

How do the potential threats from AI technology compare with those from other computer science technologies, and to bio-, nano-, and nuclear technologies?

Biological and nuclear technologies provide mush more immediate threats of weapons, yielded either by states or by small groups. Nanotechnlogy threatens to produce rapidly reproducing threats, either as weapons or accidentally, but the feasibility of this technology is still quite hypothetical. As discussed in the text and in the previous exercise, computer technology such as centralized databases, network-attached cameras, and GPS-guided weapons seem to pose a more serious portfolio of threats than AI technology, at least as of today.

Exercise 28.FAIR

Your company is working on a new model to sort people into a predicted class, $\hat{y} = \langle 0, 1 \rangle$. In the table shown, you can see the results of the model over a population of 20 samples of the true classes, $y = \langle 0, 1 \rangle$, for people across two groups, a and b. Notice the mislabelled points in each entry.

	$\hat{y} = 0$	$\hat{y} = 1$
Group a	$[0, 0, 0, 0, 0, 0, 1]$	$[1, 1, 0]$
Group b	$[0, 0, 0, 0, 1]$	$[1, 1, 1, 1, 0]$

a. Is the model **well calibrated**? If not, how might that be achieved with these data?

b. Does the model achieve **equal outcome**? If not, how might that be achieved with these data?

c. Does the model achieve **equal opportunity**? If not, how might that be achieved with these data?

d. Does the model have **equal impact**? If not, how might that be achieved with these data?

e. What constitutes a sensitive group? What if the groups, $\langle a, b \rangle$, were a person's race, such as, in the United States, black and white? What if the groups were a person's height, tall and short?

f. What constitutes a sensitive classification? What if this model was used to decide whether to give a longer sentence given predicted recidivism? What about if the model was used to decide whether to display an ad for shoes given predicted efficacy of the ad?

g. What other concerns, besides those mentioned, might arise in such attempts at classification?

h. What recommendations might you make to this company with regard to algorithmic fairness? Choose one recommendation and expand to describe how it might be implemented.

a. Yes, the model is well calibrated; the proportion for each estimated label is equal to proportion of the true label: $P(\hat{y} = 0 \,|\, a) = P(y = 0 \,|\, a) = \frac{7}{10}$, $P(\hat{y} = 0 \,|\, b) = P(y = 0 \,|\, b) = \frac{5}{10}$, $P(\hat{y} = 1 \,|\, a) = P(y = 1 \,|\, a) = \frac{3}{10}$, and $P(\hat{y} = 1 \,|\, b) = P(y = 1 \,|\, b) = \frac{5}{10}$.

b. The model does not achieve equal outcome because $P(\hat{y} \,|\, a) \neq P(\hat{y} \,|\, b)$. These data might be classified by an alternative model to meet this condition as in the following table. Nonetheless, notice how this condition compromises whether the model is well balance and does not necessarily entail equal opportunity, although it can.

	$\hat{y} = 0$	$\hat{y} = 1$
Group a	$[0, 0, 0, 0, 0, 0, 0, 1]$	$[1, 1]$
Group b	$[0, 0, 0, 0, 0, 1, 1, 1]$	$[1, 1]$

c. The model does not achieve equal opportunity because the proportion of false positives is not equal between the groups: $P(\hat{y} = 1 \,|\, a, y = 0) = \frac{1}{3} \neq P(\hat{y} = 1 \,|\, b, y = 0) = \frac{1}{5}$.

Given that the base rates between the classes are different, $P(y = 0 \,|\, a) \neq P(y = 0 \,|\, b)$, and we assume some degree of mislabeling on the part of the model, $P(\hat{y} = 0) \neq P(y = 0)$, equal opportunity is not possible to achieve unless we compromise the calibration of the model (Kleinberg, 2016). If we were willing to make such a compromise, an example classification output of these data using a new model might be as follows. Notice that we ensuingly compromise the false negative rate as well, while in the first model $P(\hat{y} = 0 \,|\, b, y = 1) = \frac{1}{5}$, now $P(\hat{y} = 0 \,|\, b, y = 1) = \frac{3}{7}$.

	$\hat{y} = 0$	$\hat{y} = 1$
Group a	$[0, 0, 0, 0, 0, 0, 1]$	$[1, 1, 0]$
Group b	$[0, 0, 0, 0, 1, 1, 1]$	$[1, 1, 0]$

d. In this case given the lack of information about the classification itself, it is impossible to say whether the model achieves equal impact, that being an extension of equal opportunity with respect to the expected utility of ensuing events caused by a classification. A sufficient answer might explain that or even formulate the costs and benefits of the classifications, $\hat{y}$, into expected utilities as Chouldechova (2017) does for the disparate impact of criminal risk sentencing.

e. Being fairly open-ended and value based, this question and the next might befit a classroom discussion. What constitutes a sensitive group is of great social, legal, and political importance. Depending on the country such groups will often have some degree of legal protection, such as race, color, national origin, sex, or religion in United States

discrimination law. This is not always the case as is apparent with current and quite recent discussions of sexual orientation, gender, ableness, and more.

f. Nonetheless, some applications of classification will demand much greater attention than others, perhaps as related to the expected impact of a classification. Clearly criminal risk sentencing demands greater attention that predictive advertising, but, as mentioned in the text, the scale of the latter as well as particular uses (such as for housing or employment) will necessitate attention as well.

g. It is a basic principle of statistics that data as well as the models based on them are biased; the observer effect tells us that as soon as we measure (decide upon some paradigmatic measure into which we project the world) we introduce bias into our data. Furthermore, see the discussion in the text with regard to whether a defendant has committed a crime as opposed to whether they have been convicted.

h. A suitable answer might expand on any of the bullet points identified in fairness and bias section. For example, using the above exercise on a real data set and model might constitute, "Examine your data for prejudice..." Alternatively to "Understand how any human annotation of data is done" one might need to systematically (by hand) go through the annotation system and classifications used, looking for those that stand out such as Crawford (2019) do with regard racist and sexist classifications in ImageNet.

Exercise 28.3REV

Concerns regarding the effect of technology on the future of work are not new. In a 1964 memo to President Lyndon B. Johnson, the Ad Hoc Committee of the Triple Revolution (1964), one comprised of notable social activists, scientists and technologists, warned of revolutions in social justice, automation, and weapons development. Like Karl Marx, John Maynard Keynes, and countless thinkers today, the committee realized that the cybernation revolution, "one brought about by the combination of the computer and the automated self-regulating machine," could free people from work, but that it faced many challenges in doing so.

> There is no question that cybernation does increase the potential for the provision of funds to neglected public sectors. Nor is there any question that cybernation would make possible the abolition of poverty at home and abroad. But the industrial system does not possess any adequate mechanisms to permit these potentials to become realities. The industrial system was designed to produce an ever-increasing quantity of goods as efficiently as possible, and it was assumed that the distribution of the power to purchase these goods would occur almost automatically. The continuance of the income-through-jobs link as the only major mechanism for distributing effective demand—for granting the right to consume—now acts as the main brake on the almost unlimited capacity of a cybernated productive system.

a. Using evidence from the text, do the quoted arguments ring true today?

b. Relate this passage to the question raised in the text about what to do with regard to automation, "do we want to focus on *cutting cost*, and thus see job loss as a positive; or do we want to focus on *improving quality*, making life better for the worker and the customer?" How does the answer to the previous question change based on which

stakeholder (e.g. a business owner, a worker, a consumer, a policy maker, etc.) is answering it? In the end, who does make those decisions around automation?

a. A suitable response here would include an historical discussion of **technological unemployment**, the magnification of **income inequality**, and relevant cautions regarding general ignorance of the **pace of change** of those effects. Certainly one can find these arguments repeated today, but what we do not know is whether the arguments are mainly accurate and have not been addressed or whether the arguments are mainly inaccurate and nevertheless have much cachet. Students will necessarily differ in their attributions of the causes: the degree to which a given technology or set of technologies is responsible for an ensuing job loss or inequality. A more nuanced answer will address the degree to which current (and preceding) technologies on the whole do not strictly displace workers (although some do), but rather change the nature of work (one's agency over it, latency of communication, location, and other trends related to globalization). Consider: What do we gain in such an accelerating economy? More happiness? More free time? More toys? These, and other, questions are raised by Wajcman (2014).

b. and **c.** Clearly the position of a stakeholder will change the degree to which they are interested in cutting costs or improving quality; in a simple consideration, a business owner wishes to reduce cost while an employee wishes to increase the quality of their life. Given the state of the world, we might then reasonably presume that it is business owners and their ilk who are the ones making decisions around automation, and workplaces in general. But, then, do workers and consumers have no say? Are they not those who elect the policy makers (in democratic countries) who create the rules for businesses? Such complications, while well warranted, expand the scope of this question to political science—there have certainly been many movements, some with spottier records than others, to give workers control over the means of production. Nonetheless, it appears reasonable to assume that most people, regardless of their position, if asked this question, would assert that they are more interested in improving the quality of their life rather than cutting cost (which could be construed as an instrumental goal for the latter). As much psychology research shows us: we do not usually act in our best interests as utility theory purports we ought. It is evident that many people do not make decisions that would be seen to increase their quality of life.

Exercise 28.NRGY

The text describes some of the **unintended side effects** which might result from artificial intelligence technologies, perhaps from **externalities** in the objective function of a system. Here consider one potential externality: energy use (and, by proxy, greenhouse gas emissions) associated with training deep learning models.

Strubell et al. (2019) make the case that modelers ought to pay more attention to the demands of training, citing Amodei (2018) as indicating how popular deep learning models increased the amount of compute time required by some 300,000 fold in between 2012 and

2018, a trend that appears to continue at the time of this writing. Indeed, Strubell et al. estimate that training a single big transformer model, the state of the art for many natural language processing tasks, would use 192 lbs CO_2, about a tenth of the emissions released by a passenger flying from New York to San Fransisco and back. They estimate that training the well-known BERT model would take 1438 lbs CO_2, a bit under that round trip cost for a passenger on a cross-country flight. Both of these figures grow thousands of times larger when neural architecture search, tuning, and other hyperparameter experiments are run.

a. Why might deep learning models be growing by such a scale? (Namely, training operations are doubling every 3.4 months as opposed to the every 2 years which used to describe the rate of change of Moore's law Amodei (2018).)

b. The cited calculations are rife with uncertainty. Where do you think that uncertainty arises?

c. What might be done to change the objective function of these models (or of the community of modelers) to incorporate the costs of energy?

a. The period which Amodei (2018) use to begin the increase in the scale of deep learning models is the emergence of AlexNet, the first major demonstration of the efficacy of deep learning approaches to the broader community. Now, as has previously been demonstrated for high dimensional data, the accuracy of such neural network approaches tends to increase as their capacity increases and as training time increases exponentially as noted in Chapter 22. While the costs of training such models is certainly immense, one presumes that the expected utility of such models is even greater (think of the competition between and the resources of today's technology giants). Thus, because of the increased accuracy of such models and the awareness of such accuracy, it is not surprising that more compute time is being invested into training them. Also see the discussion of the emergence of deep learning in the historical notes of that chapter.

b. Uncertainty arises because of a number of factors. Most modelers do not report any sort of proxy which could be used to compare compute times and those proxies which are reported are quite vague and are machine dependent. For example, computing a single floating point operation on one machine versus another may result in quite different uses of electricity and the grid system in which a machine sits largely determines the degree to which the electricity used actually results in more greenhouse gases. Lacoste (2019) breaks down the average green house gas emissions for different server regions, finding that the average carbon output of servers in Australia is more than twice that of North America.

c. There have been a few recent attempts to raise awareness of the exponentially increasing costs of training through appeals to the research community, such as those already cited as well as Bender (2021). Others, such as Henderson (2020) and Lacoste (2019) advocate for new training benchmarks, software packages to more easily measure compute time, and the like, measures which could feasibly be specified directly in the objective function of some an algorithm as a negative cost or constraint on model capacity or training epochs. Others, like Hamerly (2019), propose alternative and less energy-

intensive computational substrates, such as light.

Exercise 28.THRT

Analyze the potential threats from AI technology to society. What threats are most serious, and how might they be combated? How do they compare to the potential benefits?

It is hard to give a definitive answer to this question, but it can provoke some interesting essays. Many of the threats are actually problems of computer technology or industrial society in general, with some components that can be magnified by AI—examples include loss of privacy to surveillance, the concentration of power and wealth in the hands of the most powerful, and the magnification of particular societal biases. As discussed in the text, the prospect of robots taking over the world does not appear to be a serious threat in the foreseeable future.

Exercise 28.SOCI

Compare the social impact of artificial intelligence in the last fifty years with the social impact of the introduction of electric appliances and the internal combustion engine in the fifty years between 1890 and 1940.

The impact of AI has thus far been extremely small, by comparison. In fact, the social impact of *all* technological advances between 1970 and 2020 has been considerably smaller than the technological advances between 1890 and 1940. The common idea that we live in a world where technological change advances ever more rapidly is out-dated. It is nonetheless the case that, relative to contemporaneous innovations, artificial intelligence is having quite a large impact.

Exercise 28.CYBG

Recall the discussion of **transhumanism** from the text. Similar considerations have been made with regard to cyborgs, cybernetic organisms. Either individually or in a discussion group, consider:

a. What is a cyborg? (Perhaps look up cybernetic, too.)

b. Are you a cyborg? Do you know any cyborgs? (For example, consider existing smartphones, prosthetic devices, SCUBA gear, social media, and technologies both new and ancient.)

c. Now, do you think your great-great-grandparents would agree with your answer to the previous question? What about your great-great-grandchildren?

a. Definitions of cyborg (or human vs. transhuman for that matter) are far ranging. The aim here is to challenge students on their assumptions. Does a piece of technology have to be fully integrated with a human for them to be designated a cyborg? Answers may

include a discussion of necessity (can the modified organism survive without the modification? What about current medical technology, then? Does having a single life-saving surgery make an organism a cyborg? Over what time scale must a modification apply?), prevalence (what proportion of organisms—staying aware of what population we purport to sample from—have any given modification?), composition (does the cybernetic system have to be deterministic, command and control-like as in early considerations of AI? What about inferential systems? Must its parts be inorganic? Is a genetically-modified organism a cyborg? Must the modification be controlled using neurological signals? Why not a remote?), origin (does the origin of a modification affect the designation of cyborg? What if the part is made by evolution? What about if it is made by a person? What about evolutionary processes and people, like evolutionary algorithms?), effect (does the modification enhance or detract from an aspect of the organism, such as fitness?), and more.

b. This question and question (c) recapitulate the previous question meant to encourage students to ground their answers in examples. One might extend this exercise by challenging students' definitions with further examples—perhaps adversarially by looking for edge cases. If a student does not think that they are a cyborg from their current frame of reference, might someone in a dramatically different frame of reference disagree? To what degree is our definition of cyborg (and human!) a construction of what is not yet the case (must a cyborg always be something not yet actualized)?