Chapter 7: On Trustworthy and Ethical AI

1. Self-driving cars have the potential to vastly improve our lives. Automated vehicles could substantially reduce the millions of annual deaths and injuries due to auto accidents, many of them caused by intoxicated or distracted drivers. In addition, automated vehicles would allow their human passengers to be productive rather than idle during commute times. These vehicles also have the potential to be more energy efficient than cars with human drivers and will be a godsend for blind or handicapped people who can't drive. But all this will come to pass only if we humans are willing to trust these vehicles with our lives? **Do you think that you might be willing to trust your life to these vehicles? Why, or why not?**

I am already fearful of human drivers in general. Autonomous vehicles do not alleviate that fear whatsoever, as the self-driving artificial intelligence is ultimately programmed by humans, of whom are error-prone by nature. For instance, the Boeing 737 MAX software failure that resulted in a fatal plane crash is an example that comes to mind when considering that software is not exempt from human error. At the end of the day, the software is only as good as its human programmer, and I do not trust humans when it comes to operating heavy machinery. It should be noted that my already present fears pertaining to transportation could be clouding my judgment – after all, planes are statistically safer than cars and I still illogically prefer traveling in a car over a plane, so perhaps my viewpoint is not representative of the general human public.

2. MM enumerates a number of huge benefits that AI systems already bring to society. **Please list a few of these.**

Melanie Mitchell writes, "Current AI technology is central to services you yourself might use all the time, sometimes without even knowing that AI is involved, including speech transcription, GPS navigation and trip planning, email spam filters, language translation, credit-card fraud alerts, book and music recommendations, protection against computer viruses, and optimizing energy usage in buildings," (118).

3. MM suggests that in the near future, AI applications will likely be widespread in health care. **Please list a few of the AI applications that she foresees.**

Melanie Mitchell suggests that AI will be used for health diagnoses, drug discovery, and elderly health care (via monitoring) (119).

4. What, according to Demis Hassabis, the cofounder of Google's DeepMind group, is the most important potential benefit of AI?

Demis Hassabis believes that AI will be significant in helping humans solve complex problems that otherwise could not be solved within a reasonable amount of time. These problems include the likenesses of "climate change, population growth and demographic change, and ecological and food science," (119).

5. In discussing the phenomenon of AI taking over jobs that humans do at this point in time, MM raises the question of whether or not this will actually benefit society. In considering the question, she lists a number of jobs that technology automated long ago, suggesting that AI may simply be extending the same area of progress: improving life for humans by increasingly automating the necessary jobs that no one wants to do. **Please list a few of the jobs that technology automated long ago.**

Technology is used in assembly line robots, which is menial labor. The space industry also uses robots for exploration (119-120). Looking farther back, technology has replaced the need for the following: "..clothes washer; rickshaw driver; elevator operator; *punkawallah*…and a [human] computer," (120). In the future, AI could be used for the transportation of goods and humans, agricultural harvests, and cleaning up hazardous waste, to name a few (120).

6. What was the AI researcher Andrew NG suggesting when he optimistically proclaimed, "AI is the new electricity."

Andrew NG suggests that the potential impact of AI is ubiquitous in nature, meaning that it will have a substantial impact across a plethora of industries, just like electricity did. Mitchell clarifies that it is "...the idea that soon AI will be as necessary—and as invisible—in our electronic devices as electricity itself," (120).

7. What major difference does MM observe between electricity and AI?

The primary difference between electricity and AI that Mitchell notes is its predictability; AI is more unpredictable in its behavior than electricity (120).

8. What is "the great AI tradeoff?"

The "great AI tradeoff" is the question of how much society should rely on AI in our technology, given its duality of potential benefits and drawbacks. While AI automation may result in

improvements for humans, its tumultuous behavior and security challenges present a simultaneous threat to human beings (120-121).

9. TRUE/FALSE - Machine intelligence presents a knotty array of ethical issues, and discussions related to the ethics of AI and big data have filled several books.
10. List a couple of "positives" relating to face recognition systems.

Melanie Mitchell writes, "Face-recognition technology has many potential upsides, including helping people search through their photo collections, enabling users with vision impairments to identify the people they encounter, locating missing children or criminal fugitives by scanning photos and videos for their faces, and detecting identity theft," (122).

11. Present-day face-recognition systems have been shown to have a significantly higher error rate on people of color than on white people. Describe the ACLU study that strikingly underscored this point.

The ACLU study used Amazon's Rekognition to match faces of the 535 members of the United States Congress to a criminal database; 28 of those members wrongly matched, and 21% of the wrongful matches were African American congressional members (123).

12. TRUE/FALSE - Given the risk of AI technologies, many practitioners of AI are in favor of some kind of regulation. But simply leaving regulation up to AI practitioners would be as unwise as leaving it solely up to government agencies. The problems surrounding AI - trustworthiness, explainability, bias, vulnerability to attack, and morality of use - are social and political issues as much as they are technical ones. Thus, it is essential that the discussion around these issues include people with different perspectives and backgrounds.
13. True/False questions are often used to assess student knowledge. If a student responds with the sanctioned answer, it is assumed that they possess the sanctioned knowledge. **Please suggest an alternative use for True/False questions.**

True/False questions do not guarantee that a student knows the answer; to do that, the student should have to provide an explanation with the True/False question answer. True/False questions can be used as a view into the student perspective on an issue, depending on how the question is phrased. Additionally, these True/False questions act as a figurative highlighter of book information, making students pay closer attention to the phrases included in the questions of this nature.

14. In one example of the complexity of crafting regulations for AI systems, in 2018 the European Parliament enacted a regulation on AI that some have called the "right to explanation." This regulation requires, in the case of "automated decision making," "meaningful information about the logic involved" in any decision that affects an EU citizen. This information is required to be communicated "in a concise, transparent, intelligible and easily accessible form, using clear and plain language." *Does this regulation prohibit the use of hard-to-explain deep-learning methods in making decisions that affect individuals (such as loans and face recognition)?* Such uncertainties will no doubt ensure gainful employment for policy makers and lawyers for a long time to come. **What do you think about the highlighted question? Please say a thing or two of significance about the question.**

The highlighted question exudes the ambiguity that the EU law is shrouded in, which is never beneficial when it comes to matters of legality. Essentially, the ambiguous language has opened up the floor for loopholes, where deep-learning can be presented as "explainable" in legal terms. This clouded language does not aid progress in the proper regulation of AI (125).

15. TRUE/FALSE - The infrastructure for regulating AI is just beginning to be formed. In the United States, state governments are starting to look into creating regulations, such as those for face recognition or self-driving vehicles. However, for the most part, the universities and the companies that create AI systems have been left to regulate themselves.
16. One of the stumbling blocks in regulating AI is that there is no general agreement in the field on what the priorities for developing regulation and ethics should be. At least some attention should probably be focused on:
    a. Algorithms that can explain their reasoning.
    b. Data privacy.
    c. The robustness of AI systems to malicious attacks.
    d. Bias in AI systems.
    e. The potential "existential risk" from superintelligent AI.

    MM states her own opinion that too much attention has been given to the risks of superintelligent AI and far too little to deep learning's lack of reliability and transparency and its vulnerability to attacks. But I would like for you to venture your opinion on

prioritizing the consideration of issues surrounding AI. How would you prioritize the focus of attention on these five issues? **Please provide a list of all five elements, ordered from that which you believe is the most pressing for consideration to that which you believe is least pressing for consideration.**

- The robustness of AI systems to malicious attacks
- Algorithms that can explain their reasoning
- Bias in AI systems
- Data privacy
- The potential "existential risk" from superintelligent AI

Cyber attacks are the most concerning presently for me since AI is used in a vast array of critical infrastructure. However, I believe that if we can focus on and pivot to "algorithms that can explain their reasoning," we may be able to mitigate issues pertaining to cyber attacks and bias in AI. This, of course, involves switching systems to use explainable AI, which may not be tangible. The true existential threat in that list is the use of AI as attack vector by malicious actors, not superintelligent AI.

17. MM poses the question: If we are going to give decision-making autonomy to face-recognition systems, self-driving cars, elder-care robots, or even robotic soldiers, don't we need to give these machines the same ability to deal with ethical and moral questions that we humans have? **What do you think?**

I think this is a great idea in theory, but hard to put into practice considering there is not a singular path to being "moral." If there was, philosophy would not have multiple schools of thought. In creating a "moral machine," we also have to ask ourselves which philosophy of morality we subscribe to, and does this vary by situation? This gets convoluted very quickly in my opinion. Combined with the fact that humans do not always think logically/morally/ethically, it is difficult to gauge how "human" we should program these machines.

18. What are Asimov's three "fundamental Rules of Robotics"?

    Asimov's three "fundamental Rules of Robotics" are:

    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 the 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," (126).

19. What was Asimov's purpose in proposing the three fundamental Rules of Robotics.

Asimov's goal was to show that abiding by a set of rules like these in regard to machines is doomed to fail. Mitchell writes, "Asimov was prescient: as we've seen, the problem of incomplete rules and unintended consequences has hamstrung all approaches to rule-based AI intelligence; moral reasoning is no different," (126).

20. In Arthur C. Clarke's 1968 book *2001: A Space Odyssey*, the artificially intelligent computer HAL is programmed to always be truthful to humans, but at the same time to withhold the truth from human astronauts about the actual purpose of their space mission. HAL, unlike Asimov's clueless robot, suffers from the psychological pain of this cognitive dissonance: "He was … aware of the conflict that was slowly destroying his integrity - the conflict between truth, and concealment of truth." The result is a computer "neurosis" that turns HAL into a killer. **Please suggest one significant similarity between HAL and the AI Chatbots that are now being unleashed on the world, and one significant difference between HAL and the AI Chatbots that are now being unleashed on the world.**

Both Chatbots and HAL deal with human interaction and are programmed to respond a certain way depending on what is said by the human to the machine. However, HAL and AI Chatbots differ in the sense that Chatbots do not have a cognitive dissonance crisis at hand; they merely are programmed to respond to a human in a reasonable way with no overlying goal. On the other hand, HAL was programmed to both be truthful and withhold information from the humans conducting a mission.

21. TRUE/FALSE - The trolley problem has become a kind of symbol for asking about how we should program self-driving cars to make moral decisions on their own.

22. TRUE/FALSE - In one survey, 76 percent of the participants answered that it would be morally preferable for a self-driving car to sacrifice one passenger rather than killing ten pedestrians. But when asked if they would buy a self-driving car programmed to sacrifice its passengers in order to save a much larger number of pedestrians, the overwhelming majority of survey takers responded that they themselves would not buy such a car. According to the authors, "We found that participants in six Amazon Mechanical Turk

studies approved of utilitarian AVs (that is, autonomous vehicles that sacrifice their passengers for the greater good) and would like others to but them, but they would themselves prefer to ride in AVs that protect their passengers at all costs."

23. TRUE/FALSE - A prerequisite to trustworthy moral reasoning is general common sense, which is missing in even the best of today's AI systems.