Reading Assignment from Melanie Mitchell's "Artificial Intelligence: A Guide for Thinking Humans"

Assignment by Carrie Corcoran

Chapter 6: 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?

Self-driving vehicle technology would have to notably improve before I would even consider using one. At this time, the number of crashes involving self-driving vehicles is entirely too high to suggest that this technology is safe.

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

Some examples she lists are speech transcription, GPS navigation, and email spam filters.

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.

She foresees AI assisting in diagnosing diseases, discovering new drugs, and home monitoring of patients.

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

He believes the most important benefit is in data analysis and scientific modeling, as this can be applied to big problems such as climate change and food science.

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 are 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.

Among the jobs she lists are clothes washer, lift (elevator) operator, and computer (a person who did complex mathematical calculations by hand).

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

He was saying that AI will transform every industry in the next several years the way electricity did when it was introduced.

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

The difference is that electricity was widely understood before it was used widely, whereas AI has noted transparency problems.

8. What is "the great AI tradeoff?"

The Great AI tradeoff refers to whether we should allow extensive use of AI to enrich our lives despite the risks, or should we be more cautious and avoid the potential problems caused by AI.

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.

True

10. List a couple of "positives" relating to face recognition systems. List a couple of "positives" relating to face recognition systems.

Some of the positives include being able to find missing children, and helping visually impaired people identify others. The negatives include loss of privacy and unequal accuracy across racial groups, meaning that people of color are more likely to be misidentified.

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 used Amazon’s facial recognition system to test pictures of 535 members of congress, to see if they matched anyone in a criminal database. The system incorrectly matched 28 members, disproportionately people of color.

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.

True

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.

Other forms of questions, such as multiple choice or short answer questions, are alternatives to True/False questions.

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 the4 “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.” This opens the floodgates for interpretation. What counts as “meaningful information” or “the logic involved”? 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.

This regulation appears to prohibit deep-learning methods until such time as deep-learning methods that can explain their logic are developed. This may be a difficult regulation to follow and may prevent some AI technology being used in the EU, but this might be an acceptable trade for EU citizens wanting a greater degree of control over their technology.

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.

True

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 focussed on:
• Algorithms that can explain their reasoning.
• Data privacy.
• The robustness of AI systems to malicious attacks.
• Bias in AI systems.
• 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 believe is the most pressing for consideration to that which you believe is least pressing for consideration.

- Data privacy
- Bias in AI systems
- Algorithms that can explain their reasoning
- The robustness of AI systems to malicious attacks
- The potential "existential risk" of 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 solders, 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 would agree that any system making life-or-death decisions should be able to weigh the full consequences of those decisions, and that includes the ethical questions involved.

18. What are Azimov's three "fundamental Rules 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 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.

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

His purpose was to show how such a set of rules would inevitably fail.

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.

HAL and AI Chatbots are both rule-based AI systems. One major difference between them is that HAL was programmed to be truthful to humans, while AI chatbots such as ChatGPT can freely fabricate its responses.

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.

True

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."

True

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

True