Kuncheng Feng
CSC 466

Chapter 7: On Trustworthy and Ethical AI

→ Reading/Mining/Discussion Assignment

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'm not willing to trust my life to these vehicles, I don't think reliability of such vehicles have been proven yet, and I'm afraid that in case of emergency, it will deem that sacrificing me is the best moral choice for society.

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

AI systems have already been used in 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.

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.

AI can assist physicians in diagnosing diseases, suggest treatments, discovering new drugs, and monitoring the health and safety of elders in their homes. AI can also assist in scientific modeling and data analysis.

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

There are complex problems that even the smartest set of humans can't make an advancement on during their lifetime, AI is the assistance and the solution to that.

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.

A few boring, exhausting, degrading, exploitative or downright dangerous jobs have been listed: Cloth washer; rickshaw driver; lift operator; pukah-wallah (manual fan waver); computer (a human that performs tedious calculations).

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

AI will soon be as necessary and as invisible in our electronic devices as electricity itself.

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

Electricity was well understood before it was implemented, but right now we do not have much understanding of AI.

8). What is “the great AI tradeoff?”

“Should we embrace the abilities of AI systems, which can improve our lives and even help save lives, and allow these systems to be employed ever more extensively? Or should we be more cautious, given current AI’s unpredictable errors, susceptibility to bias, vulnerability to hacking and lack of transparency in decision-making?”

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 “negatives” relating to face recognition systems.

Positive: Help search through photo collections, people identification for visually impaired, locate missing children or criminal fugitives, detect identity theft.

Negative: Violates privacy, assigns value to people and sells it as a service, wrongly identifies people as criminals.

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.

They tested Amazon’s Rekognition system on 535 members of the US Congress, 28 matched to be criminals, and it has a higher rate for black congressmens.

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.

Have them explain their logic, how they arrive at such conclusions.

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.

I think it is a good regulation, and I just assume it means that AI should be able to explain things in a way that 95% of the population can understand. I also think this will push further understanding of AI, because if it can't explain things very well, probability because it doesn't understand it very well (Standard expectation for students).

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.

In my opinion, since this is a free market economy, companies should just aim for what they think will make them the most money.

But if I’m an AI developer, I would aim to do so in the following manner:

- Algorithms that can explain their reasoning.
- The robustness of AI systems to malicious attacks.
- Bias in AI systems.
- Data privacy.
- The potential “existential risk” from 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 don’t think we need to give these machines that same ability to deal with ethical and moral questions, as it significantly raises the bar for AI developments, making it difficult for new AI companies to start their own project. However I do think that it would be nice if an AI can recognize a morally challenging situation and follow a predefined solution.

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.

To demonstrate what such a set of rules would inevitably fail, trapping the robot in an endless loop.

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.

- Similarity: The AI Chatbots are always polite to humans, never in a way to verbally abuse humans.
- Difference: The AI Chatbots do not have a self conscious, nor the power to control our surroundings.

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 systems.

True