Module 1: Critical Thinking

AI Safety

Nolan Byrnes

CSC510 – Foundations of Artificial Intelligence

Colorado State University – Global Campus

Professor Bingdong Li

May 22nd, 2023

Module 1: Critical Thinking

Artificial Intelligence (AI) Safety is an important topic that we all need to consider as the field of AI grows. At surface level, regulation sounds like an easy way to help improve AI safety, however, there are some challenges that prevent regulation from occurring, and may stifle innovation when implemented. This paper discusses some of the challenges that we face in regards to AI Safety and regulation, suggestions on how we can do our best to keep the AI solutions that are released to the public are safe.

Amazon created a hiring tool that used artificial intelligence that would search for the best job candidates. Dastin claims that “Amazon’s computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period”. The data that the model was being trained on were mostly from males, so Amazon’s AI solution favored male over female applicants. The issue was not Amazon’s AI solution itself, but the data that was used to train the AI model. If biased data is being used to train AI, the AI solution would end up having biased results in its decision making.

I think that most would agree that we would all want to have AI solutions that are fair and free of bias, and at surface level, having regulation in place can achieve this goal. The challenge is that “There is no single definition of fairness that will apply equally well to different applications of AI” and “Given the many complex sources of unfairness, it is not possible to fully ‘debias’ a system or to guarantee fairness” (Madaio et. al, n.d., pg. 1). In Microsoft’s AI Fairness checklist, they propose that we consider all stakeholders of a given AI solution, and the potential harm that the system may cause to those stakeholders, including demographic groups (Madaio et. al, n.d., pg. 3). In the example of Amazon’s recruiting AI tool, they did not purposely develop their solution to be discriminate to women, however, due to the data that was gathered, the AI model trained itself to be more in favor towards men.

If the team who were developing the recruiting tool thought about the people the AI solution would affect (job applicants), the demographics of those job applicants (men and women), and the realization that their data set was disproportionately male resumes, they could have addressed the issue earlier on in the project. Amazon did edit their program to fix this issue of gender bias, however, there was no guarantee that the program would find other ways to discriminate against other demographics. Amazon ultimately disbanded the project.

Regulation could ultimately hinder development of AI solutions. Having regulations which require the algorithms used to be explainable would be challenging since AI is a challenging topic, and the general public does not have the knowledge to understand the algorithms being developed. Furthermore, there could not be anything wrong with the algorithm, but the data set that is being used. If regulations were in place when Amazon developed their recruiting tool, then perhaps the disproportion of men and women resumes would have been caught, but then you would then need to make sure that all ethic groups are being represented equally. The data for some ethnic groups may not be available which could limit the amount of data that is being used to train the AI models.

OpenAI claims “that learning from real-world use is a critical component of creating and releasing increasingly safe AI systems over time” (OpenAI, 2023, para. 5). OpenAI “cautiously and gradually release new AI systems—with substantial safeguards in place—to a steadily broadening group of people and make continuous improvements based on the lessons we learn” (OpenAI, 2023, para. 6). With OpenAI’s approach, although it is a frightening thought to have the general public as guinea pigs for AI solutions, OpenAI is testing their solutions for 6 months prior to release and make continuous improvements to the software post release. I do think that this approach is rational, as with ChatGPT, it would be impossible to test every combination of inputs and review every response, and they have thoroughly tested the solution for half a year before release.

Like OpenAI, I also agree that regulations should be put in place, however, I am not convinced that regulations can be put in place in a timely manner. Technology and artificial intelligence are improving at a rapid pace, and by the time it takes for politicians to put regulations in place, damage may have already been done. For example, The General Data Protection Regulation (GDPR) “officially went into effect in 2016, but European Union member states had until May 25, 2018 to fully implement GDPR policies” (Bell, 2022, para. 4). With the rapid advancement of technology, a lot can happen in a two year period, and damage can be done before any regulations can take effect.

Programming ethics into AI solutions would be difficult to achieve due to ethics being difficult to define. Ethical decisions can be situational, and it also depends on the culture on what is considered to be ethical. MIT researchers designed an experiment called the moral machine, which crowdsourced “people’s decisions on how self-driving cars should prioritize lives in different variations of the ‘trolley problem’” (Hao, 2018, para. 1). What they found is “that participants from individualistic cultures, like the UK and US, placed a stronger emphasis on sparing more lives”, whereas other cultures would prefer to spare the individual in the car (Hao, 2018, para. 7). It would be difficult to program ethics into AI solutions because of how ethics are viewed in different cultures.

To keep AI solutions safe, companies who are implementing software solutions should have a code of ethics, follow it throughout the development process, and make it publicly available. By having their code of ethics publicly available, it allows the general public to know where they stand where it comes to AI. The thought is to apply pressure to approach AI ethics in a way that would keep all their users happy.

Elon Musk signed an open letter “calling for a six-month pause on the technology’s development because of its ‘profound risks to society’” (Metz et. al., 2023, para. 2). At the same time, Musk is also ramping up his own AI activities by hiring top AI researchers from Google’s DeepMind at Twitter, and speaking publicly “about creating a rival to ChatGPT that generates politically charged material without restrictions” (Metz et. al., 2023, para. 4). It is a bit hypocritical, especially since he is calling for pause on AI development while mentioning that he is creating a rival to ChatGPT with less restrictions, which could pose a greater threat than what OpenAI is currently doing with ChatGPT. I get the impression that Elon Musk is calling for a pause in development so that he can catch up with the technology.

I do not agree with having a pause in the development of AI technology because it does not seem to be logical to pause innovation, but we should be careful to release the developed AI technologies to the public. Snapchat released its own AI Chatbot called ‘My AI’, which “may not always be accurate, and its responses "may include biased, incorrect, harmful, or misleading content" (Gerkin, 2023, para. 1). Although the new feature may be useful and entertaining to some, it is in a way a social experiment. I think that although there is a lot of benefit of AI, forcing it upon the public as quickly as possible could have negative unintended consequences.

Conclusion

As the field of artificial intelligence grows, we need to be sure that we do so carefully and in an ethical manner. Ethics are difficult to program within AI solutions since all cultures are different, and what may be ethical for one culture, may not be for another. Some of the issues that may appear with AI solutions could be due to the data that is being provided to train the AI models, which may not always be evenly distributed. Regulations on AI developments such as having AI algorithms explainable, as the algorithms could be too complicated for the general public to understand, and some of them not completely understood by the developers themselves.

Applying regulations to use distributed data sets that cover every gender and demographic could also hinder development, since it would limit how much data could be used due to smaller demographic groups. To keep AI safe, companies that use AI should keep their code of ethics publicly available, to apply pressure for companies to use practices that would be favorable to their users. AI solutions should also be carefully released to the public until we understand the negative consequences we would face with the release of it. It seems that we are in an AI race, as companies want to have the best AI solutions, but in doing so, it could be damaging to society unless we really think about the potential negative consequences of AI solutions on society as much as the positive.

**REFERENCES**

Bell, M. (2022, April 5). When Did GDPR Go Into Effect? SixFifty. [www.sixfifty.com/blog/when-did-gdpr-go-into-effect/](http://www.sixfifty.com/blog/when-did-gdpr-go-into-effect/)

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. U.S. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

Gerkin, T. (2023, April 26). Snapchat introduces AI chatbot to mixed reviews. BBC News. <https://www.bbc.com/news/technology-65388258>

Hao, K. (2018, October 24). Should a self-driving car kill the baby or the Grandma? Depends on where you’re from. MIT Technology Review. <https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/>

Madaio, M., Stark, L., Vaughan, J. W., & Wallach, H. (2020, March). AI fairness checklist. Microsoft Research. <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t6dA>

Metz, C., Mac, R., & Conger, K. (2023, April 27). Elon Musk ramps up A.I. Efforts, even as he warns of dangers. The New York Times - Breaking News, US News, World News and Videos. <https://www.nytimes.com/2023/04/27/technology/elon-musk-ai-openai.html>

Madaio, M., Stark, L., Vaughan, W., & Wallach, H. (n.d.). AI Fairness Checklist. Microsoft. <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t6dA>

OpenAI. (2023, April 5). Our approach to AI safety. <https://openai.com/blog/our-approach-to-ai-safety>