Kerrigan de Meij
Data/AI Ethics & Law: Plan of Action

Areas of Machine Learning & AI that can lead to Bias and Discrimination

1. Many Large Language Models (LLMs) have been trained on supposed openly accessible user-generated sites (ex. Reddit) however, structural barriers cause marginalized communities to be less welcomed to contribute to these sites. Thus, many LLMs are trained on datasets where under-represented populations' data and viewpoints have less of an impact, and the LLM serves those who already have the most privilege in society. Many Political, Healthcare, and Financial Systems across the globe already facilitate inequality among various groups of people. Society cannot allow technology to become an addition to this list as well.

2. Technology has been promised to connect people, however the majority of AI usage and its budget has been towards the surveillance of people. The growth in budgets for immigration and border policing has gone from $1.2B in 1990 to $25.2B in 2019. Smart borders and other means of surveilling technology has divided societies through harming those populations who are victims of political violence, climate change, economic inequality, etc. while serving the viewpoints and desires of those already privileged with a developed government and relative economic stability.

3. Many employees face discrimination and inequalities regarding hiring processes without the involvement of technology and AI, and this process being automated has shown in many ways to worsen the situation. EEOC Chair Charlotte Burrows says, specifically, that it's a warning signal when employers aren't ready to explain in a clear and understandable way the tools by which applicants are being screened out through AI during the hiring process.

4. A significant theme throughout ethical Data & AI discussion is that the size of the training dataset does not guarantee diversity. Many solutions when examining discrimination in ML models and datasets involve simply collecting more data in order to better diversify the dataset. However, sometimes only considering the dataset and what it contains is not sufficient enough to truly reduce its discriminatory tendencies. There may be a root problem with how the dataset and model is being deployed, rather than just the content of the dataset.

Steps to Reduce Bias and Discrimination in Data & AI

1. All datasets embed a particular view of the world, meaning that realistically there's no such thing as a completely unbiased dataset. However, working in a world that is already economically unstable and unequal in so many areas, Machine Learning & AI work must focus on identifying these inevitable biases and decide what viewpoint should be portrayed in the dataset and model. This depends on the purpose of the model, who will be impacted, and why it was created. Using a heterogeneous approach to consider the institution, social system, decision-making culture, etc. for which the model was built will help reduce these discriminatory outcomes of Machine Learning & AI Models.

2. Because LLMs and other Machine Learning Models' benefits and costs are clearly not distributed evenly across communities, it is critical that the researchers and developers of datasets and ML systems follow a standardized approach to maintain the integrity and privacy of a dataset and/or developed model. Although one institution or entity's purpose is to connect people, another may use that dataset and system for different needs and outcomes. In which case, the dataset is not in alignment with the goals and purposes of the new entity and should not be used with the same system and framework. This can be achieved by strictly enforcing the disclosure of the original purpose of a dataset model and framework, and regulating its accessibility to the public.

3. There is a clear-cut regulatory system that organizations must follow when interviewing potential employees, and distributing such work opportunities in an equitable way, therefore AI should also be built to coincide with these regulatory policies in order to keep a centralized, non-discriminatory hiring process across the nation. The automation of hiring processes is a clear example of the need for a standardized way to disclose evaluation methods and documentation of other data governance measures. If employers cannot clearly display how AI tools are being deployed in the hiring process, this framework of technology usage is far from where the regulatory standards should be.

4. Sometimes reducing bias and discrimination with ML and dataset work involves considering the way that the system is being deployed, such as facial recognition technology. If this work has shown to cause harm to certain communities, such as being disproportionately inaccurate when identifying specific races and genders, it may be necessary to reconsider the original purpose and intended benefit of this technology.