Undergraduate Thesis Prospectus

Developing and Deploying Robust Fairness-Aware AI Algorithms

(Technical Research Project in Computer Science)

The Struggle for Safe and Equitable Artificial Intelligence in the United States

(Sociotechnical Research Project)

by

Matthew Cahill

October 27, 2023

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Matthew Cahill

STS advisor: Peter Norton, Department of Engineering and Society

General Research Problem

How may AI systems be implemented for optimum social benefit?

The rapid integration of AI systems in society underscores their transformative potential. When used correctly, these systems can provide the necessary tools for significant technical and sociological advancements. Beyond this technical wonder lies a more poignant question: How can these systems be tailored for optimum societal advantage while balancing their destructive capabilities? Mismanaged AI has the capability to deliver detrimental outcomes when unleashed onto vulnerable populations. A study performed by human rights advocacy group, Amnesty International, highlights the risks of AI in surveillance. It discusses how companies such as Google and Facebook's surveillance-based business models are "inherently incompatible with the right to privacy and pose a treat to a range of other rights including freedom of opinion and expression, freedom of thought, and the right to equality and non-discrimination" (Beneson, 2019). The developers of AI systems must take into consideration every potential threat to human rights and individual freedoms prior to deployment. They must be educated on these implications and made fully aware of the disastrous effects that come with ill-informed AI development.

Developing and Deploying Robust Fairness-Aware AI Algorithms

In the deployment of AI systems, how can algorithmic biases be prevented or managed?

AI is quickly becoming intertwined with every facet of life, and software developers must be prepared to harness its power for sociological benefit, as opposed to making society's most vulnerable the victims of a colossally dangerous timebomb. The early workings of these AI algorithms, which are trained on historical data, have already been shown to propagate and even exacerbate existing societal biases. An MIT study showed that commercial AI facial recognition tools developed by major tech companies such as IBM and Microsoft had "higher error rates for darker-skinned individuals" (Buolamwini, 2018). A similar study conducted by ProPublica highlights how certain AI models developed for use in the United States criminal justice system are "biased against African American individuals" in how they work to "detect future criminals" (Angwin, 2016). Addressing potential bias in AI system development is not only an ethical imperative, but also critical in ensuring the reliability and trustworthiness of this technological feat that will undoubtedly continue to be essential in everyday living.

The technical research project goals include working through the designing the structure and thought process of developing the following: Firstly, an AI algorithm that can identify and mitigate biases in its predictions. Secondly, a monitoring system that can detect emergent biases in deployed models. Lastly, a research project that will educate stakeholders on the vital importance and methods of ensuring fairness-aware AI.

The project's unusual constraints are as follows: First, the inner workings of fairness-aware algorithms must be transparent so their decision-making processes can be analyzed and justified. Second, any method being used to detect bias must ensure that the data of individuals is kept confidential. Third, as AI interacts with every facet of life and society,

tailoring the training of algorithms to accurately take into account unique situational biases is computationally expensive.

The current methods being used to address AI fairness include adversarial training, adjusting weights in training data, and post-hoc calibration (Bellamy, 2018). Tech companies deploy tools such as Google's TensorFlow Fairness Indicators in an effort to appear as though they are actively avoiding bias and discrimination in their systems. In reality, the minimal solutions that Google and other big tech AI development companies deploy are reactive, meaning they address biases once identified. While tools such as TensorFlow have a variety of fairness metrics, there exists little to no guidance regarding which of these metrics are optimal. This project will address how to better combat these biases in the form of prevention and limitation rather than reaction. Structuring AI algorithmic training such that it has built-in mechanisms and strategies to eliminate biases is the foundation of this technical project. The contribution of this project's findings will work to better the sociological landscape through minimizing the damage that historically disadvantaged and vulnerable populations receive from untrustworthy AI systems.

This project's scope will not involve producing a full-fledged AI model that accurately and efficiently removes all bias and discriminatory practices from its decision-making processes. Rather, it will focus on the theoretical framework behind developing the system and the conceptual inner workings of the algorithm training. This begins with a comprehensive review of fairness models and the current shortcomings of AI system decision-making with respect to specific demographics of historically targeted populations. The framework will cover data collection, model development, real-time monitoring and adjustment practices, and stakeholder education.

By the project's end, a tangible framework that will fully cover the necessary steps to achieve the complete development and deployment of robust fairness-aware AI algorithms will have been produced. The impact of this research project will improve the general trust in AI systems as they will be deployed and operated equitably. Next, this newfound AI framework will actually be developed, trained, and deployed to further promote a safer AI future.

The Struggle for Safe and Equitable Artificial Intelligence in the US

In the US, how are equity advocates striving to achieve regulations that limit inequities in the application of AI systems?

In the United States, equity advocates seek to prevent discriminatory bias, limit surveillance, protect privacy, and ensure human mediation in the training and deployment of AI systems. As artificially intelligent technology advances, understanding the efforts of equity advocates to combat these concerns is crucial to solving the problem. Data collected by the American Civil Liberties Union suggests that unchecked AI could exacerbate societal disparities, disproportionately impacting marginalized communities. An ACLU study highlights how AI systems used by certain property companies to evaluate potential tenants rely solely on court records and other datasets that "have their own built-in biases that reflect systemic racism, sexism, and ableism" (ACLU, 2021).

In recent years, a plethora of research has been conducted relating to bias within AI systems. A study concerning economic customer discrimination analyzed bookings through home-sharing company Airbnb. The study found that a significant number of hosts were rejecting customers based on "race, age, gender, and other factors" (Murphy, 2016). Customer rejections were based on AI scans of their online public profiles on social media websites. The case concluded that Airbnb's AI algorithms resulted in "incorrect predictions on profits and an impact on people through racial discrimination" (Murphy, 2016). During the height of COVID-19, the American Medical Informatics Association analyzed how AI bias contributed to the disparities gap. Although medical professionals hoped the newly touted AI advancements would help guide treatment decisions during the crisis, it was impacted by bias, and ultimately failed to "proactively develop comprehensive mitigation strategies during the COVID-19 pandemic" which resulted in "exacerbating existing health disparities" (Rice, 2020).

As the world's leading AI company, Microsoft exemplifies how tech companies take advantage of technological algorithmic advancements to cause societal harm, while operating under the guise of "committing to the practice of responsible AI by design, guided by a core set of principles" (Microsoft, 2023). OpenAI is a leading AI development company that actively perpetuates "the unequal treatment of demographic groups through content moderation" (Rozado, 2023). The ACLU is "calling on the Biden administration to take concrete steps to bring civil rights and equity to the forefront of its AI and technology policies" as AI has already begun to deepen racial and economic inequities (ACLU, 2021). In response to the ACLU's call, the Biden administration has "published a landmark Blueprint for an AI Bill of Rights to safeguard Americans' rights and safety, and U.S. government agencies have ramped up their efforts to protect Americans from the risks posed by AI" (White House, 2023). The American Civil Liberties Union, the Biden Administration, Microsoft, and OpenAI will all serve as participant groups. Microsoft and OpenAI manipulate artificial intelligence to help their companies profit off of society's most vulnerable and historically disadvantaged populations, while the ACLU and Biden Administration work to combat this systematic manipulation.

To address the numerous societal vulnerabilities that come with the rapid expansion of AI technology, industry experts, policymakers, and academics must work to anticipate how to develop and use AI. This inevitably will result in AI raising new ethical, legal, and governance issues, including racial discrimination, gender bias, and issues related to customer awareness of AI's role in decision outcomes (Varsha, 2023). These conversations are the necessary first step in mitigating the risks associated with this technology. As AI development is in its early stages, minimizing vulnerabilities is still possible.

References

Akselrod, O. (2023, July 3). *How Artificial Intelligence Can Deepen Racial and Economic Inequities*. American Civil Liberties Union. www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities

Angwin, J. (2016, May 23). *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bellamy, R. (2018, Oct. 3) *AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.* IBM. arxiv.org/pdf/1810.01943.pdf

Beneson, P. (2019, Nov. 21). *Surveillance Giants: How the business model of google and facebook threatens human rights*. Amnesty International. www.amnesty.org/en/documents/pol30/1404/2019/en

Buolamwini, J. (2018, Jan. 15). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. Proceedings of Machine Learning Research. proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Microsoft. (2023, March 8). *What is Microsoft's Approach to AI?* Microsoft. news.microsoft.com/source/features/ai/microsoft-approach-to-ai

Murphy, L.W. (2016, Sep. 8). *Airbnb's Work to Fight Discrimination and Build Inclusion.* Laura Murphy & Associates. cdn.geekwire.com/wp-content/uploads

Rice, B. (2020, Aug. 17). Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19. Journal of the American Medical Informatics Association. doi.org/10.1093/jamia/ocaa210

Rozado, D. (2023, Feb. 2). *The Unequal Treatment of Demographic Groups by CHATGPT/OpenAI Content Moderation System*. Substack. davidrozado.substack.com/p/openaicms

The United States Government. (2023, Sep. 12). *Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI*. The White House. www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration

Varsha, P.S. (2023, April 1). *How can we manage biases in artificial intelligence systems - A systematic literature review.* International Journal of Information Management Data Insights. doi.org/10.1016/j.jjimei.2023.100165