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

An STS Research Paper
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Matthew Cahill

March 14, 2024

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

The Struggle for Safe and Equitable Artificial Intelligence in the US

Equity advocates in the United States seek to prevent discriminatory bias, limit surveillance, protect privacy, and ensure human mediation in the training and deployment of AI systems. Unchecked, AI could exacerbate social disparities, disadvantaging marginalized communities.

In a study of Airbnb, Murphy (2016) found that many hosts refused customers because of their "race, age, gender, and other factors." Some hosts used AI scans of guests' online public profiles on social media websites. Murphy concluded that Airbnb's AI algorithms yielded "incorrect predictions on profits," in effect constituting "racial discrimination" (2016). During the COVID-19 pandemic, the American Medical Informatics Association studied how AI bias contributed to healthcare disparities. Medical professionals had hoped that AI systems would improve treatment decisions, they found encoded biases. Because agencies failed to "proactively develop comprehensive mitigation strategies during the COVID-19 pandemic," in using AI systems they were "exacerbating existing health disparities" (Rice, 2020). According to an ACLU study (Akselrod, 2021), AI systems that evaluate potential tenants rely on court records and other datasets that "have their own built-in biases that reflect systemic racism, sexism, and ableism."

Although Microsoft, the leading AI company, develops AI systems that cause social harm, it claims it has been "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 (Akselrod, 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). As AI development is in its early stages, minimizing vulnerabilities is still possible. Despite efforts by groups like the Biden administration to combat systemic bias, equity advocates still face challenges in advocating for effective regulations. The influence of the tech industry and complexity behind understanding the makeup of AI systems make it difficult for lawmakers to create and enact policies. In response to this, a comprehensive learning course must be undertaken by every future lawmaker/policymaker in the United States. Artificial intelligence is here to stay and will only continue to advance at a rapid rate. The only way to effectively prevent bias in AI is through properly educating our representatives on the makeup of artificially intelligent technology and the training of the algorithms behind them.

While United States policymakers are beginning to discuss at length “the risk AI technologies carry in making discriminatory decisions,” few understand the algorithmic makeup

and language models that serve as this technology's innerworkings (Neill, 2023). Real accountability can only be achieved when entities are held responsible for their decisions. Policymakers must suggest governance requirements for the development and deployment of AI. The best course of action to take at this moment is to provide "AI assurance". This term can loosely be defined as "a range of AI accountability processes and tools… that can support this process by proving that an AI system is legal, effective, ethical, safe, and otherwise trustworthy" (US Department of Commerce, 2023). Only when our government officials are experts in the realm of artificial intelligence and machine learning, can we begin to combat inherent algorithmic biases. The questions that must be directed at developing AI systems by those battling these issues at the intersection of technology, race, and prejudice are as follows: "What does this product do? What is its daily use? Who does it disempower?" (Basiouny, 2023). This last question is the most important. Once we take into consideration who is disproportionately impacted by the algorithms we create, we can effectively address them in a way which stops potential harm. Promoting diversity and inclusion in the development of artificial intelligence is critical to ensuring fair and equitable AI systems. Boston Consulting Group conducted a study that found companies with "more diverse management teams have 19% higher innovation revenues" (Lorenzo, 2018). Not only does diverse leadership encourage new ideas, innovations, and business benefits in the AI realm, but also promotes cultural and general life perspectives that are currently being ignored by AI developers and the algorithms they craft.

Although generative AI has the potential to "widen the racial economic gap in the United States by $43 billion each year" moving forward, its power could actually be harnessed to remove economic mobility barriers for marginalized communities when deployed thoughtfully (McKinsey, 2023). New wealth created by digital and AI capabilities sees the median Black

household obtain roughly 15 percent of the wealth held by the median White household. With annual global wealth creation from generative AI projected to be about $7 trillion at the current technological development trajectory, roughly $2 trillion of said pool is expected to enter the United States economy directly given the country's share of global GDP. This new wealth rooted in generative AI value creation suggests "the United States could gain nearly $500 billion in household wealth" (McKinsey, 2023). This wealth will inevitably be split up inequitably if the previously mentioned distribution of wealth remains unchanged. Black Americans bring in only about 38 cents of every dollar of new household wealth despite representing over 13 percent of the United States population. As this trend continues, the racially disparate distribution of new wealth created by the development and deployment of generative AI could "increase the wealth gap between Black and White households by $43 billion annually" by 2045 (McKinsey, 2023).

Deploying generative AI with an equity lens is in the hands of the leaders of this and the coming generations. These leaders can either use it in a manner that promotes fairness, opportunity, and inclusion, or they can idly watch it grow uncontrollably before their eyes as it furthers dangerous inequity and disastrous societal implications. Reskilling workers is a vital step in the process to approach generative AI in the right way. As opposed to role-specific-only training, workers in marginalized communities should turn their focus to "foundational and nuanced skill building" with the help of funding and resources provided by equity advocates (McKinsey, 2023). Another critical step is ensuring democratized access. Generative AI tools should be available and equally accessible to all. AI development companies should not offer exclusive access or benefits to privileged communities, as this exacerbates inequity and endangers vulnerable populations. Whether this comes in the form of government regulation, or

an understanding between AI development companies based on the desire for AI equity from educated stakeholders, restrictions must be put in place.

While the rapid expansion of generative AI into mainstream applications has been seemingly seamless, it comes with a blatant disregard of robust safety measures. Currently, most model developers lack the requirement for significant safety features when third parties deploy the generative AI models they develop through API. While the Federal Trade Commission's (FTC) $5 billion fine of Meta was a step in the right direction towards holding companies accountable for their failure to adequately secure the private information and data of users, there is still significant ground to cover on the front of inequity prevention in AI. Legal liability is society's primary tool of ensuring accountability and responsibility when laws are violated. With this in mind, lawmakers must create laws that will serve as a deterrent for AI developers either unintentionally, or intentionally, causing harm in their deployment of inequitable AI technologies. The Center for American Progress is currently exploring "the topic of liability for developers and deployers of AI models" to effectively "recommend legislative solutions" (Shahi, 2024). This kind of work is essential to aiding the necessary effort to establish and enforce comprehensive governance structures for generative AI.

At their best, developers must aim to safeguard their systems, be fully transparent with users and stakeholders, and uphold responsibility for the technological tools they create and deploy. Unfortunately, the AI development community in the United States is currently lacking on all fronts. As almost every major generative AI developer in the United States offers third parties the ability to deploy LLM technology to their own use cases through APIs, new AI systems are being deployed at an unprecedented rate (Shahi, 2024). As generative AI tools are being deployed at this rate, perfect safeguards are virtually impossible to be put into place

meaning security vulnerabilities and data breaches are occurring at an alarming rate. Although major AI development companies such as OpenAI, Microsoft, Meta, and Google all have general use policies that warn against using their tools for harmful activities such as generating malware or planning activities that could lead to death or bodily harm, there exist no details or information on any of their sites relating to what "exactly constitutes a violation of these usage policies" (Shahi, 2024). With little to no information regarding policy violation repercussions associated with any of these major companies, a dangerous gray area is created that leaves room for the exploitation of vulnerable communities.

Balancing technological innovation with adequate safeguards and regulations is one of the most significant challenges facing society as it integrates with AI. A study conducted by Just Security has found that Chatbot GPT is "already being leveraged by criminals to perpetrate fraud against unsuspecting victims" (Garrison, 2023). Concerns have been raised surrounding face recognition and the AI trained algorithms behind them being deployed in criminal investigations. The Georgetown University School of Law's Center for Privacy and Technology released a report that highlights the fact that jurisdictions often "lack the proper policies and procedures necessary to govern the use of face recognition evidence, and that has led to rights violations and wrongful arrests" (Garrison, 2023). Research has shown that law enforcement in nearly every US state has begun to rely heavily on facial recognition technology. This technology used by police departments across the country has been shown to misidentify "women and people of color more often than white men" (Johnson, 2022). Even with this information, most of the United States has criminal justice systems in place where neither police nor prosecutors are legally obligated to inform those accused of crimes if facial recognition was part of the investigation that led to their arrest.

There currently exists no federal law in the United States that governs AI development or use. Though AI regulation is failing at the federal level, certain states have moved to regulate the new technology in promising ways. Alabama, California, Colorado, Connecticut, Illinois, Louisiana, New Jersey, New York, North Dakota, Texas, Vermont, and Washington, are 12 states that have "enacted laws that delegate research obligations to government… entities to increase institutional knowledge of AI and better understand its possible consequences" (Lerude, 2023). Senate Majority Leader Chuck Schumer assembled a majority of the United States Senate, CEOs of major technology companies, as well as labor and civil rights leaders for a timely AI Insights Forum. The participants of the forum all overwhelmingly agreed that "AI regulation is necessary" (Lerude, 2023).

The exacerbation of existing forms of societal discrimination will continue to take place until proper AI development guardrails are put into place. Certain states such as California, Colorado, Connecticut, Massachusetts, New Jersey, and Rhode Island, are "proposing and passing legislation to ensure that the adoption of AI does not perpetuate bias against protected classes" (Lerude, 2023). The District of Columbia has taken a critical step in relation to this subject as their lawmakers have proposed legislation that would "prohibit entities that use algorithms to make service eligibility determinations from making determinations based on protected characteristics, subject to civil action by the attorney general and individuals suffering relevant discrimination" (Lerude, 2023). Without safeguards like this in place, a biased and inequitable AI algorithm can unleash all kinds of harm against those who do not have the resources necessary to defend themselves.

The use of AI in hiring practices has also begun to be regulated in certain states such as New York. In 2021, New York City lawmakers passed a law that "restricts the use of automated

decision systems in the screening of candidates by requiring employers to conduct bias audits, publish results, and notify candidates of the use of such tools, subject to civil penalty" (Lerude, 2023). Discriminatory biases in hiring practices play a major role in the perpetuation of economic inequality in the United States. There exist automated decision systems that are inherently built on bias and prejudice in their algorithmic inner workings. In certain places around the country other than New York City, flawed hiring practices are being used to bar those in marginalized communities from landing a role they are qualified for.

Combating racial bias in AI-powered technologies is necessary at all levels of the education system. Nonprofit equity advocate organizations such as Digital Promise as well as Edtech Equity Project have created a new product certification called "Prioritizing Racial Equity in AI Design" (Harold, 2022). The product certification is being applied to both companies and educational institutions to evaluate if the creators of education technology tools are taking the necessary steps to "identify and question their own biases and assumptions, ensure the data used to train their artificial intelligence systems aren't tainted by bias, and provide educators and families alike with more visibility into how their products actually work and the risks they might pose" (Harold, 2022). Building tools, systems, and solutions to work towards solving the racial inequity problems that face school systems across the country aim to help those living at margins who are often ignored during the design and development phases of the creation process. Equity advocates in the United States should follow the example of organizations such as Digital Promise and Edtech Equity Project. Meaningful change in the realm of AI bias prevention cannot take shape without the contributions of city-based organizations that focus on assisting one specific area in need.

With all of the problems that AI can present in relation to racial bias and economic disparities in school systems, it can be used to transform school systems for the better when used correctly. Training AI algorithms in a way that is not predatory and looking to mine personal data, but rather, focusing on how to provide K-12 students with a "personalized learning experience tailored to their individual preferences and needs, immediate feedback on their work and answers to their questions, and increased access to tutoring and other educational materials" is the key to correcting this pressing issue (Diebold, 2022). An example of the application of AI in education administration is identifying situations for administrative intervention for at-risk students. Studies have shown that "more than 50 percent of high schools in the United States use predictive algorithms as an early warning system to identify students at risk of dropping out" (Diebold, 2022). AI systems being used in this manner not only have the ability to flag certain students showing signs of academic struggle, but can also help create appropriate and personalized intervention tactics to ultimately help a given student succeed. Most current systems identify students as on or off track based on preset thresholds and how their grades intersect with said thresholds. New systems have begun to be developed based on machine learning models to track different predictive indicators of a student in a marginalized community nearing dropout of an educational institution. One of these new systems is the Chicago Early Warning Indicator, which "focuses on ninth grade students, their accumulated credits, and low or failing grades" to catch vulnerable and marginalized students in time before giving up their educational ambitions (Diebold, 2022). When used properly and bias is effectively mitigated in the algorithmic development process, AI can be applied as a tool to benefit those in marginalized communities.

Society has already begun to feel the serious harm that comes with a lack of delivery in the development of AI systems. As machine learning models are trained using historical data and expected to give unbiased, optimal results, a few questions must be posed regarding the gathering and treatment of said data prior to its use in the model: “Where are you getting this data from? Have you cleaned the data? Have you looked at the data to see if it reflects current trends? Was there diversity when you first collected the data, or is this based on your own bias?” (Snell, 2022). As AI technology evolves, United States citizens must choose: either stand idly by as unregulated AI harms marginalized populations and society’s most vulnerable or work with equity advocates to ensure artificial intelligence benefits all walks of life.

The struggle for safe and equitable AI in the United States demands collaboration among equity advocates, policymakers, and the technology industry. By addressing discriminatory biases, surveillance, privacy, and the need for human mediation in artificial intelligence, we can create systems that benefit not only society’s most privileged but the underprivileged as well.

References

Akselrod, O. (2021, July 13). *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

Basiouny, A. (2023, Nov. 10). *Diversity Is Critical for the Future of AI.* Knowledge at Wharton. knowledge.wharton.upenn.edu/article/diversity-is-critical-for-the-future-of-ai

Diebold, G. (2022, April 1). *How AI Can Improve K-12 Education in the United States.* Center for Data Innovation. www2.datainnovation.org/2022-ai-education.pdf

Garrison, B. (2023, Jan. 11). *Regulating Artificial Intelligence Requires Balancing Rights, Innovation.* Just Security. www.justsecurity.org/84724/regulating-artificial-intelligence

Harold, B. (2022, April 12). *Why Schools Need to Talk About Racial Bias in AI-Powered Technologies.* EducationWeek. www.edweek.org/leadership/why-schools-need-to-talk-about-racial-bias-in-ai

Johnson, K. (2022, Mar. 7). *How Wrongful Arrests Based on AI Derailed 3 Men's Lives.* Wired. www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/

Lerude, B. (2023, Nov. 1). States Take the Lead on Regulating Artificial Intelligence. Brennan Center for Justice. www.brennancenter.org/our-work/research-reports/states-take-lead-regulating-artificial-intelligence

Lorenzo, R. (2018, Jan. 23). *How Diverse Leadership Teams Boost Innovation.* BCG. www.bcg.com/publications/2018/how-diverse-leadership-teams-boost-innovation

McKinsey. (2023, Dec. 19). *The impact of generative AI on Black communities.* McKinsey Institute for Black Economic Mobility. www.mckinsey.com/bem/our-insights/the-impact-of-generative-ai-on-black-communities

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

Neill, B. (2023, Sep. 21). *Eight AI-related US policy issues for boards and management to consider.* Ernst & Young. www.ey.com/en_us/public-policy/ai-policy-landscape

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

Shahi, M. (2024, Feb. 1). *Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone.* Center for American Progress. www.americanprogress.org/article/generative-ai-should-be-developed-and-deployed-responsibly-at-every-level-for-everyone/

Snell, K. (2022, Feb. 13). Lack of diversity in AI development causes serious real-life harm for people of color. NPR. www.npr.org/2022/02/13/1080464162/lack-of-diversity-in-ai-development-causes-serious-real-life-harm-for-people-of-

United States Department of Commerce. (2023, April 13). *AI Accountability Policy Request for Comment.* National Telecommunications and Information Administration. www.federalregister.gov/documents/2023/04/13/2023-07776/ai-accountability-policy-request-for-comment

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

White House. (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