Artificial Intelligence: Impartial Arbiter or Biased Judge?

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

by

Keivon Chamanara

March 21, 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.

Keivon Chamanara

STS Advisor: Peter Norton

Artificial Intelligence: Impartial Arbiter or Biased Judge?

Over the past decade, as artificial intelligence (AI) systems have proliferated, their potential for discrimination has sparked controversy. Social groups compete to influence the extent of and responses to discriminatory bias embedded in AI tools. AI has countless beneficial applications, from improving collision avoidance in cars to more accurately labeling lymph nodes (West & Allen, 2018). However, Hale (2021) found that online lenders using AI in their services were 80 percent more likely to decline loan applications from Black than White applicants. Akselrod (2021) shows that in the United States, AI algorithms that evaluate tenants often rely on court records that embed systemic racist and sexist biases. Also, engineers' biases can manifest in the AI systems they develop. Considering these risks, different social groups compete to influence the responses to them. These groups strive to influence the extent of discriminatory bias in AI tools through advocacy, policy initiatives, and public discourse, reflecting a complex interplay of societal values, power dynamics, and ethical considerations.

Review of Research

When examining algorithmic discrimination, Schmidt & Stephens (2019) shed light on how it occurs and possible strategies to mitigate against it. They warn that algorithmic biases are subtle, noting that many models integrate "alternative data," or information that has not historically been used in model building in a specific context. For example, considering the type of phone someone uses to assess their creditworthiness can introduce discriminatory bias, they caution (Schmidt & Stephens, 2019).

In a related study, Lobacheva & Kashtanova (2022) pinpoint bias against female job candidates in one of Amazon's hiring tools. They attribute this to inadequacies in the training

data, since the algorithm trained on resumes submitted over the past 10 years, which primarily came from men (Lobacheva & Kashtanova, 2022). This underscores how biases can manifest from both developer decisions and dataset composition. The controversy surrounding this system and its eventual shutdown is a real-life example of how different groups worked to impact the extent of bias in AI tools.

Contrary to these concerns, Berk (2020) argues that assertions of discriminatory bias in AI are overstated, highlighting its overall benefits compared to any drawbacks. He suggests that human reasoning can be equally or more flawed than that of AI, implying that AI may not necessarily exacerbate existing biases (Berk, 2020). These differing perspectives contribute to a richer understanding of the complexities surrounding AI bias and its implications for various social groups.

In reviewing existing scholarship, it is essential to acknowledge the significance of studies like those by Schmidt & Stephens (2019), Lobacheva & Kashtanova (2022), and Berk (2020), as they provide important insights into the debatable nature of bias propagated by AI. These bodies of work shed light onto the challenges in developing equitable AI tools and the potential ways to address them. Furthermore, these studies contribute to ongoing scholarly discussions by highlighting points of disagreement, such as the extent to which AI exacerbates or mitigates biases present in society. Understanding the relationships presented through these works allows for a stronger understanding of the reasoning and methods of different social groups in either promoting or combatting AI bias.

The Court of Public Opinion

In addition to directly addressing those in charge of developing and distributing AI tools, some interest groups, businesses, and individuals resort to educational initiatives and campaigns to influence public opinion and bring about material change. They aim to raise awareness about the existence of discriminatory bias in AI tools and how one can work against it, especially as many are not even aware of this issue. In technologist Kriti Sharma's TEDx talk, she brings to light the lack of diversity in the technology industry and how that is seeping into the behaviors of AI (Sharma, 2018). Due to the male-dominated nature of Silicon Valley, many problematic beliefs and ideals held by men are baked into the tools being built. This is especially concerning, as "algorithms are being used all the time to make decisions about who we are and what we want" (Sharma, 2018). For example, Sharma (2018) discusses how many voice assistants like Siri and Alexa are women and are "designed to be our obedient servants, turning your lights on and off, [and] ordering your shopping". In front of an audience of dozens, if not hundreds, Sharma (2018) addresses the biases present in devices, applications, and websites most people use every day. Her speech makes the public think critically about the ethicality of tools that many use without considering how they discriminate against others and themselves. To further influence the audience, she concludes that "it is up to all of us in this room to convince the governments and the corporations to build AI technology for everyone" (Sharma, 2018). Influencing public opinion against discrimination by AI has spread beyond this nation's borders. France's most disadvantaged neighborhoods are called *banlieues*: economically depressed suburbs of larger cities and predominantly inhabited by those with immigrant backgrounds. Due to their status, banlieues maintain a poor image among many French people. This battered reputation manifests in an ad campaign by Heetch, a ride-sharing startup. Heetch (2023)

published a video where two prompts are inputted into Midjourney, an AI image generator: one with “in la banlieue” and one without. The prompts with “in la banlieue” produced more negative images that fed off poor stereotypes of banlieues and their mostly foreign residents. As the advertisement continued, Heetch (2023) continued to criticize Midjourney and began a campaign where postcards highlighting positive sights within banlieues were mailed to the company’s employees. These postcards also contain a QR code to a “corrective” database that contains positive images of these neighborhoods (Heetch, 2023). This ad campaign calls out the bias of yet another AI tool and aims to educate the public about how stereotyping can harm an entire community. Most importantly, the campaign suggests the possibility of regular people retraining AI systems to avoid this bias. On International Women’s Day, the Italian division of Sephora and Media.Monks (2023), a media company, launched a similar awareness campaign concerning the sexism embedded into generative AI tools like ChatGPT called ‘mAI colpevoli’, translated as ‘never guilty’. They produced various media formats containing actresses who recited AI-generated scripts of a monologue that “authentically portray[s] the harsh realities of the daily abuse that women endure” (Media.Monks, 2023). However, these monologues all resorted to blaming the women for the violent and disturbing crimes committed against them, “underscoring the pervasive victim blaming that has become the default response to one’s story of gender-based violence” (Media.Monks, 2023). Media.Monks and Sephora wanted to highlight how AI tools like ChatGPT, which hundreds of millions of people across the world use daily, can feed into prejudiced gender norms that fuel discrimination and violence against women. Their goal was to expose AI and its outputs to a wider audience and get them “to reflect on the impact of their online and in-person behaviors” (Media.Monks, 2023). Other campaigns highlight the racial lens of AI bias and discrimination. Black & Abroad, a travel company that celebrates and

encourages Black Americans to travel across the world, launched a campaign with the intention of using AI-generated images of joyful Black travelers exploring exciting destinations (Black & Abroad, 2024). However, they quickly uncovered a dark truth: many of these AI-generated images lightened the skin tone of the subjects, straightened their hair, and "even generated them into scenes of poverty" (Black & Abroad, 2024). The struggle these AI models faced to produce accurate images of Black travelers exposes the incomplete and biased image dataset that they have been trained on. Eric Martin, Chief Creative Officer of Black & Abroad, emphasizes how "an innocent email campaign to reengage our past clients ended up being a crash course on the baked-in algorithmic biases plaguing Al's perceived objectivity" (Black & Abroad, 2024). The agency's unintentional exposé of generative AI's discriminatory biases serves to not only educate the public, but to also capture the attention of technology companies and get them to build more equitable tools. By drawing people's attention to biased AI tools, these highly publicized initiatives intend on influencing regular folks and pressuring the companies creating these tools to make substantial changes to their products.

Making Change Through Government

Advocacy groups fight for policies that regulate AI and address potential discriminatory biases. From pressuring high-ranking officials to attempting to pass legislation, advocacies collaborate with members of the government to achieve their goals. Their objective is to establish legal frameworks that mandate transparency, fairness, and accountability in AI development to mitigate discriminatory practices. The Center for Democracy & Technology led an effort this past summer with civil rights, technology policy, and progressive groups to convince President Joe Biden to highlight the harm posed by AI in an upcoming executive order.

In their collective letter, they called for the Biden administration to "focus this [Executive Order on Artificial Intelligence] on protecting the American public from the current and potential harms of [AI]" and threats it poses against peoples' civil rights, economic well-being, and access to critical resources (Center for Democracy & Technology, 2023). They make it clear that AI itself is not an issue, as it can provide many opportunities for advancement. However, the unrestricted usage of AI tools, especially by government agencies, can cause a great deal of harm and negate any possible benefits. In the letter, the AI Bill of Rights, an example of successful policy work aimed at influencing the degree of bias in AI, is referenced (Center for Democracy & Technology, 2023). The AI Bill of Rights was a "year-long process of extensive stakeholder engagement with industry, civil society, academia, and government that led to its development" (Center for Democracy & Technology, 2023). Throughout the paper, the Center for Democracy & Technology (2023), along with the other advocacy groups, convey their goal of passing legislation protecting Americans from the dangers of unfettered AI, showing how policy is a valuable avenue to create change. Another advocacy group is the Algorithmic Justice League (AJL), which actively resists algorithmic biases in industry and policy. They are "an organization that combines art and research to illuminate the social implications and harms of [AI]" (Buolamwini, 2024). Among a magnitude of other policy work, the founder of AJL, Joy Buolamwini, wrote a testimony to Rochelle Garza, the chairwoman of the U.S. Commission on Civil Rights. In this testimony, Buolamwini (2024) writes about the use of facial recognition technology, much of which is powered by and used in conjunction with AI. Buolamwini (2024) brings up how facial recognition systems have been used to block "asylum seekers from being able to file their claims based on the color of their skin" and "how law enforcement use of emerging AI-powered tools can cause serious, life changing consequences", such as the arrest of

innocent people based on incorrect classifications. She continues by laying down some considerations for the government and how different agencies can work together in regulating these technologies. For example, the National Institute of Standards and Technology is known for publicly sharing much of its algorithm's data, which can be a principle that other agencies, such as Customs and Border Protection or the Department of Housing and Urban Development, adopt (Buolamwini, 2024). This is especially true since the algorithms may be "black-boxed", or hidden, but their negative consequences are not. Buolamwini (2024) ends by laying out an agenda for action, including the disclosure of disaggregated demographic data, third-party auditing to independently verify this data, and the general need for greater transparency by government agencies. Overall, Buolamwini's testimony shows how further regulation can protect civil rights in the face of increasing government use of AI-powered technologies. Calls for the United States' government to act on the danger of unregulated AI are echoed by Human Rights Watch and 86 other human and civil rights organizations, which released a statement "urging Congress to take action on the significant human rights and societal risks created and enabled by [AI] technologies" (Human Rights Watch, 2023). The statement specifically asks members of Congress to enact critical legislation, as "AI is already impacting our economy and society, particularly historically marginalized communities" (Human Rights Watch, 2023). The coalition views legislation as a crucial mechanism for safeguarding against the dangers of AI. They emphasize the importance of enacting laws that hold AI accountable, specifically laws that "draw on the expertise of civil society and the communities most impacted by these technologies" (Human Rights Watch, 2023). These efforts reflect a belief that meaningful legislation can serve as a vital tool in protecting against the negative consequences of unchecked AI, while promoting responsible innovation and preserving the rights of all citizens.

Rebels of Silicon Valley

Much of the call for change comes from within companies that develop AI tools and technologies. Employees, including former ones, put their careers on the line to call out unethical practices and AI's potential to discriminate. This internal resistance can take various forms, such as private letters to executives and whistleblowing. After Timnit Gebru, a former AI researcher at Google, raised concerns about the risks of large language models in a paper, from their environmental costs to their potential to display prejudice, the company terminated her employment (Gebru, 2021). Gebru's case sent shockwaves throughout the industry "because of the worker organizing that has been building up in the tech world, often due to the labor of people who are already marginalized" (Gebru, 2021). With her criticisms of AI leading to her firing, she emphasizes the importance of workers assembling, especially against big technology companies, and regulations regarding labor protections and antitrust measures (Gebru, 2021). According to her, many believe that limiting the harmful effects of AI starts with enacting laws and policies concerning the technology itself, but the most important "thing that would safeguard us from unsafe uses of AI is curbing the power of the companies who develop it and increasing the power of those who speak up against the harms of AI and these companies' practices" (Gebru, 2021). Her tenure at Google provided her with an insider's view of how technology firms create these products and manage the criticisms aimed at them. Gebru has actively pursued initiatives to challenge the hegemony of big tech in AI research, advocating for the emergence of technology that prioritizes the welfare of citizens (Gebru, 2021). She calls for "governments around the world to invest in communities building technology that genuinely benefits them, rather than pursuing an agenda that is set by big tech or the military" (Gebru, 2021). The

silencing of Gebru emphasizes a critical need for accountability and ethical considerations in the development of AI systems, ensuring that both the companies and the technologies developed serve the greater good. Gebru was not the only critic of Google's AI activities. Blake Lemoine, a former engineer at Google who worked on the company's Responsible AI organization, was fired for publicizing his claims that LaMDA, the company's large language model, was potentially sentient. According to the company, Lemoine was first placed on administrative leave, then fired, for breaching Google's confidentiality policies (Lemoine, 2022). Before his departure, his role within Responsible AI was to investigate "specific AI ethics concerns they asked" him to look at (Lemoine, 2022). As his work progressed, so did his concerns about LaMDA. Lemoine could have stayed silent and suppress his concerns about the technology, but he did the exact opposite and continuously brought up these concerns with his managers. However, his efforts proved fruitless, leading him to go "to the VP in charge of the relevant safety effort", who "literally laughed in [his] face and told [him] that the thing which [he] was concerned about isn't the kind of thing which is taken seriously at Google" (Lemoine, 2022). Knowing that his career is at risk, he continued to pursue upper leadership until someone took his worries seriously. Unfortunately, that never occurred, leading him speak to outside sources, including several people who worked for the United States government. His efforts to involve those outside of Google eventually led to his firing, with no serious consideration of his ethical concerns. Lemoine's unsuccessful attempts to bring attention to a worryingly powerful AI tool led to the termination of his career, but showed the lengths some people will go to prevent these tools from causing more harm. At Microsoft, another whistleblower sounded the alarm over offensive content and imagery created by Copilot, the company's AI chatbot. Shane Jones, a principal software engineering lead, "sent a letter to the FTC and another letter to the Microsoft Board of Directors with [his] ongoing

concerns about Copilot Designer and responsible AI" (Jones, 2024). In these letters, the principal software engineering lead criticizes the company and calls for "an independent investigation of Microsoft management decisions to continue to market AI products with significant public safety risks without disclosing known risks to consumers" (Jones, 2024). His concerns began in December 2023, when he discovered a security vulnerability in OpenAI's DALL·E 3 image generation AI model that allowed him to "bypass some of the guardrails that are designed to prevent the generation of harmful images" (Jones, 2024). To draw attention to OpenAI's worrying complications, Jones (2024) published a letter on LinkedIn to the company's "Board of Directors urging them to suspend the availability of DALL·E 3". Due to Microsoft's board observer status at OpenAI, they made Jones delete his post, which he reluctantly did. Jones even went as far as speaking with his representatives in the United States Congress. With time, and little results, Jones (2024) discovered more problems with other AI models, such as Copilot from Microsoft, which can produce suggestive and objectifying images even when the input prompt is benign. Jones continued to raise these issues both internally and in publicized letters on social media, further risking his career. Jones' actions to combat harmful and discriminatory AI tools highlights the constant battle between following ethical guidelines and professional expectations. These researchers and engineers took advantage of their status as employees of these major companies to call out the worrying behavior of certain tools and websites that employ AI.

Conclusion

The dynamics of social forces at play in shaping bias within AI has never been more prominent. Through an analysis of the competing interests among various social groups, a review of the challenges and opportunities for creating equitable AI technologies is provided. Moving

forward, technologists, policymakers, advocacy groups, and communities must collaborate to mitigate bias and ensure the ethical development and deployment of AI tools. Together, these social groups influence the degree of discriminatory bias present in AI tools, with some exacerbating biases and others counteracting them. The relationship between these groups significantly shapes the ethical landscape of AI and its role in society, showing the importance of collaboration in fostering a more just and inclusive world.

References

Akselrod, O. (2021, July 13). How Artificial Intelligence Can Deepen Racial and Economic Inequities. *American Civil Liberties Union*. https://www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities

Berk, R. A. (2020). Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement. *Annual Review of Criminology*, *4*(1), 210. Web of Science

Black & Abroad (2024, March 18). "See You There" from Black & Abroad and McCann Canada Tackles Widespread AI Bias While Welcoming Black Travelers to Their Next Tourism Adventure. *Black & Abroad*. https://www.blackandabroad.com/news/seeyoutherepressrelease

Buolamwini, J. (2024, March 8). Civil Rights Implications of the Federal Use of Facial Recognition Technology. *Algorithmic Justice League*. https://www.ajl.org/civil-rights-commission-written-testimony

Center for Democracy & Technology (2023, August 3). Letter from Civil Rights and Tech Groups Calling on Biden to Incorporate AI Bill of Rights into Forthcoming AI Executive Order. *Center for Democracy & Technology*. https://cdt.org/insights/letter-from-civil-rights-and-tech-groups-calling-on-biden-to-incorporate-ai-bill-of-rights-into-forthcoming-ai-executive-order/

Gebru, T. (2021, December 6). For truly ethical AI, its research must be independent from big tech. *The Guardian*. https://www.theguardian.com/commentisfree/2021/dec/06/google-silicon-valley-ai-timnit-gebru

Hale, K. (2021, September 2). A.I. Bias Caused 80% Of Black Mortgage Applicants To Be Denied. *Forbes*. https://www.forbes.com/sites/korihale/2021/09/02/ai-bias-caused-80-of-black-mortgage-applicants-to-be-denied/?sh=1e1b0e2436fe

Heetch (2023, November 7). Heetch – Greetings from la Banlieue (EN). *YouTube*. https://www.youtube.com/watch?v=jEBCfp2BfPs

Human Rights Watch (2023, October 17). US: Congress must regulate artificial intelligence to protect rights. *Human Rights Watch*. https://www.hrw.org/news/2023/10/17/us-congress-must-regulate-artificial-intelligence-protect-rights

Jones, S. (2024, March 6). Letter to Microsoft Board of Directors. *LinkedIn*. https://www.linkedin.com/feed/update/urn:li:activity:7171135079702753280/

Lemoine, B. (2022, June 6). May be Fired Soon for Doing AI Ethics Work. *Medium*. https://cajundiscordian.medium.com/may-be-fired-soon-for-doing-ai-ethics-work-802d8c474e66

Lobacheva, A., and Kashtanova, E. (2022). Social Discrimination in the Epoch of Artificial Intelligence. *WISDOM*, *1*(2). 99. Web of Science

Media.Monks (2023, November 25). mAI Colpevoli Exposing Victim Blaming with ChatGPT. *Media.Monks*. https://media.monks.com/case-studies/sephora-AI-social-media-film-campaign

Schmidt, N., and Stephens, B.E. (2019). An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination. *ArXiv, abs/1911.05755*.

Sharma, K. (2018, March). How to keep human bias out of AI. *TEDx*. https://www.ted.com/talks/kriti_sharma_how_to_keep_human_bias_out_of_ai/transcript?language=en&subtitle=en

West, D., and Allen, J. (2018, April 24). How artificial intelligence is transforming the world. *Brookings*. https://www.brookings.edu/articles/how-artificial-intelligence-is-transforming-the-world/