Elizabeth Binkina

18 April 2025

Peter Darch

Automating Bias: The Case for Socio-Technical Ethics in Artificial Intelligence

**Introduction**

As artificial intelligence (AI) continuously reshapes many automated systems such as the hiring process, urgent questions rise around bias, fairness, and accountability. This paper explores ethical implications of using historical data to train AI algorithms, especially when such data reflects historical biases and discrimination and excludes certain features as well as underrepresented demographic groups. The Amazon’s 2014 hiring algorithm case, which was ultimately discontinued due to its gender bias, examines how biased training data can lead to discriminatory outcomes and how these issues can be reduced in future AI development.

In 2014, Amazon developed an AI-powered resume screener using machine learning (ML) to streamline hiring. The system, trained on ten years of historical data, largely reflected the male-dominated tech industry. This resulted in the AI throwing out female resumes and favoring male applicants. Despite efforts to eliminate the bias, Amazon scrapped the tool in 2015 because of the difficulty in ensuring fairness in automated systems. This case has become a widely cited example of how AI systems can unintentionally perpetuate inequality when built on biased data.

The research question is: What are the ethical implications of training AI models on historical data that excludes certain variables or demographic groups? This question addresses a core challenge in AI ethics—the reality that data-driven models reflect and reinforce patterns found in their training data, even when those patterns are unjust. If left unchecked, biased models perpetuate discrimination under the guise of objectivity and efficiency. Identifying strategies to address this can help highlight the broader social consequences of deploying AI in high-stakes areas like hiring, where fairness and equal opportunity are legally and morally mandated. This paper argues that training AI hiring systems on historical data leads to discriminatory outcomes that violate core ethical principles of fairness, transparency, and accountability. While AI has the potential to reduce human bias, it can amplify systemic inequalities when not carefully designed and monitored. Developers must proactively identify and correct bias in training data, incorporate human oversight, and ensure transparency in algorithmic decision-making.

**Literature Review**

The ethical implications of training AI algorithms with biased historical data—often excluding important variables or demographic groups—are critical concerns in AI ethics. Literature highlights the consequences of biased datasets in areas like economic decisions and hiring. While consensus exists on the risks of training AI models on biased data, scholars call for more robust governance frameworks and long-term ethical accountability in AI development. This review explores themes of bias inheritance, algorithmic discrimination, and the importance of human oversight and diverse design teams.

A key concern is that AI systems inherently reflect the data they are trained on. In simple terms, if an algorithm uses biased data, the output will also reproduce and reinforce those biases. Whittaker et al. (2019) explain that “If something is missing from the data, say images of people with dark skin, these people will be missing from the AI model, and thus won’t be recognized or included” (p. 9). This exemplifies how training models on incomplete or non-representative datasets could be dangerous.

They further cite real-world examples, including an autonomous vehicle fatally striking a pedestrian because the AI lacked data on wheelchairs, bicycles, and other non-normative street scenarios. These examples emphasize a central issue: AI systems cannot learn what they are never exposed to.

Algorithmic discrimination results from biased data, where ML amplifies these biases. Varona and Suárez (2022) note that “discrimination has both an origin and cause of bias once the outcomes of today’s discriminatory decisions, based on yesterday’s biases, populate tomorrow’s datasets” (p. 2). This is particularly concerning in fields like hiring and housing, where biased outcomes can make serious detriments over time. In mortgage lending, for instance, Zou and Khern-am-nuai (2023) found that Black and Hispanic applicants were disproportionately denied loans—even when equally qualified—due to biased training data from the Home Mortgage Disclosure Act. Without intervention, these patterns risk becoming self-perpetuating.

Human intervention is critical in identifying and correcting algorithmic bias. Hall and Ellis (2023) stress the importance of including women and minorities in the development process. Their research suggests that diverse design teams are better equipped to foresee and mitigate harms to marginalized communities. They also call for structured audits and ongoing performance monitoring of AI systems, especially when deployed in contexts involving critical decision-making, like employment (p. 12). This marks an important shift in the literature—from purely technological solutions to socio-technical strategies that emphasize accountability and diversity.

While much of the literature focuses on technical fairness metrics and bias-mitigating strategies, few studies address the need for legal and institutional governance. There is a gap in understanding the long-term social consequences of excluding certain features and demographic groups from automated decision-making, as well as the need for interdisciplinary perspectives that incorporate ethical, legal, and organizational responsibility. This emphasizes the need for more ethical reflection and policy development.

Altogether, this literature provides a comprehensive foundation for examining the ethical risks of biased AI systems. It establishes that bias in training data is not merely a technical flaw, but a profound ethical concern with real-world consequences. These sources support the central claim of this paper—that AI systems trained on historically biased data will continue to replicate and amplify discrimination unless developers take inclusive, transparent, and proactive measures. They show how AI ethics must move beyond technical performance and incorporate human-centered values such as fairness, accountability, and social justice.

**Findings**

The case of Amazon’s 2014 AI hiring tool offers a powerful example of how biased training data can lead to discriminatory outcomes, even when the intention behind automation is not malicious. Amazon, aiming to increase efficiency and reduce recruiter workload, developed an AI system to evaluate and rank job applicants based on past hiring data. However, because the tool was trained on a decade of historical resumes—where the majority of applicants and hires were male—it learned to replicate and reinforce existing gender disparities.

As a result, the algorithm penalized resumes that included references to all-women’s colleges or terms like “women’s chess club,” while favoring those with typically male-coded language such as “executed” or “captured.” These exclusionary patterns were not programmed deliberately; rather, they emerged from the data itself. This case demonstrates how AI systems can perpetuate systematic inequalities when built on non-representative or biased datasets.

What makes this case especially significant is that Amazon engineers attempted to correct the bias—but the model continued to identify gender proxies, such as extracurricular activities or educational background. This persistence of bias, even after direct adjustments, highlights the limitations of purely technical fixes when deeper structural inequalities exist in the training data. It underscores the critical importance of involving diverse perspectives and ethical scrutiny early in the development process, rather than reacting to the issue after deployment.

Furthermore, the system lacked interpretability: hiring managers had little insight into how decisions were being made. The opacity undermined accountability and left applicants with no clear recourse to contest biased outcomes—violating a key principle of ethical AI. This undermines trust and creates serious legal and ethical concerns under equal employment laws. Amazon’s tool was built internally by an engineering team using proprietary data, with little external review or transparency.

Beyond gender bias, this case reveals the larger ethical dilemma posed by AI systems that rely on historical data. If the data reflects exclusion, inequality, or past injustices, then the model built upon it will almost inevitably do the same. Not only does this produce discriminatory outcomes, but it also risks institutionalizing these patterns. When flawed decisions are made by machines, they can be harder to detect and challenge than when made by humans.

Ultimately, Amazon discontinued the tool in 2015 after internal audits revealed its discriminatory effects. While this decision was ethically responsible, it was also reactive, meaning the damage had already occurred. A more proactive approach, including rigorous auditing of training data and inclusive development practices, might have identified the bias before damage was done.

This case reinforces the idea that good intentions do not eliminate ethical risks. The AI system was developed to streamline hiring, but it ended up reproducing systemic discrimination. Fairness, accountability, and transparency must be central design principles in high-stakes decision-making. Amazon assumed that technological solutions were better than human judgment but didn’t consider the ethnicity being more important. Without structural changes in data practices and development culture, AI systems will continue to mirror inequality.

**Discussion & Conclusion**

The findings from Amazon’s 2014 hiring tool highlight a broader ethical dilemma in AI development: algorithms trained on biased historical data are likely to replicate and even amplify discriminatory practices. It becomes clear that AI technology reflects the social, cultural, and institutional environments in which it is created. While AI is often used as a means to increase objectivity and efficiency in hiring, this case shows that, without careful oversight, it can exacerbate existing inequalities under the guise of automation. This ethical problem is not just a technical issue—it’s a socio-technical issue rooted in unjust histories that are reproduced through digital systems.

This reflection aligns with the literature emphasizing that AI systems inherit the biases of their creators and the data they are trained on. As Hall and Ellis (2023) argue, addressing such systemic issues requires more than just correcting flawed code—it demands a commitment to diverse development teams, frequent bias audits, and robust institutional support. The Amazon case supports this idea by showing how a lack of diverse perspectives and oversight led to the deployment of a tool that unfairly penalized female candidates. Although the company attempted to remove explicit gender markers, the system continued to identify indirect proxies—highlighting how deeply embedded bias can be. This supports Whittaker et al.’s (2019) warning that AI cannot recognize or fairly evaluate individuals it has not been taught to understand.

To address these challenges, developers and organizations must take a multi-step approach. First, they need to ensure that training data is diverse, representative, and regularly checked for bias. Second, AI algorithms have to be transparent so that decisions can be justified fairly.

Third, there must be meaningful human oversight at every stage of the development process. This includes diverse development teams as well as institutional policies that hold organizations accountable for biased and harmful outcomes. In addition to technical reforms, more cultural and policy updates are needed. Legal frameworks must evolve to account for algorithmic decision-making in hiring and other areas. Organizations must prioritize ethics alongside efficiency, and regulators must develop guidelines for fairness audits. Ethical AI is not only about avoiding harm—it is about building systems that actively promote equity and inclusion.

Moreover, these findings expose a fundamental flaw in the prevailing narrative that AI is inherently objective. The belief in algorithmic objectivity often leads to misplaced trust in automated systems, even when their decisions are biased. The Amazon case reveals that even a company with vast technical resources and expertise can overlook critical ethical issues if it fails to recognize the structural injustices embedded in its data. It also reflects the literature in its importance to shift from reactive fixes to proactive solutions. As Varona and Suárez (2022) point out, biased systems don’t just appear out of nowhere—they emerge from cycles of discrimination that become embedded in datasets and perpetuated by algorithms.

In conclusion, Amazon’s hiring tool serves as a cautionary example of what can happen when efficiency is prioritized over equality in AI systems. It illustrates that without proper education in ethics, history is bound to repeat itself, leading to harmful and unjust outcomes. To prevent such outcomes in the future, developers, organizations, and policymakers must embrace a holistic approach to AI ethics—one that prioritizes fairness, inclusivity, and accountability from the ground up. Only then can we ensure that AI enhances human decision-making without reintroducing the very biases we aim to eliminate.

**Works Cited**

Kodiyan, A. (2019, November 12). *An overview of ethical issues in using AI systems in hiring with a case study of Amazon’s AI based hiring tool*. Research Gate.  <https://www.researchgate.net/publication/337331539_An_overview_of_ethical_issues_in_using_AI_systems_in_hiring_with_a_case_study_of_Amazon's_AI_based_hiring_tool>

Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., ... & West, S. M. (2019). Disability, bias, and AI. *AI Now Institute*, *8*, 11.

<https://ainowinstitute.org/wp-content/uploads/2023/04/disabilitybiasai-2019.pdf>

Hall, P., & Ellis, D. (2023). A systematic review of socio-technical gender bias in AI algorithms. *Online Information Review*, *47*(7), 1264-1279.

<https://www.emerald.com/insight/content/doi/10.1108/oir-08-2021-0452/full/html>

Varona, D., & Suárez, J. L. (2022). Discrimination, bias, fairness, and trustworthy AI. *Applied Sciences*, *12*(12), 5826.

<https://www.mdpi.com/2076-3417/12/12/5826>

Zou, L., & Khern-am-nuai, W. (2023). AI and housing discrimination: The case of mortgage applications. *AI and Ethics*, *3*(4), 1271-1281.

<https://www.researchgate.net/publication/365372598_AI_and_housing_discrimination_the_case_of_mortgage_applications>

**Statement on AI** : I used generative AI (Microsoft Copilot and ChatGPT) for the purpose of refining and proofreading my writing, without changing my sentence structure. Based on the rough draft/ thoughts I wrote, I asked chatGPT to make an outline for me highlighting important points and to organize my ideas. I copied and pasted my writing and asked AI to proofread it for grammar, and to delete repetitive words and sentences. I also asked AI to check if my paper incorporated certain criteria for the assignment.