ALGORITHMIC BIAS AND GENDER DISCRIMINATION IN AI: THE NEED FOR LEGAL ACCOUNTABILITY

Ashifa A Saheed, School of Legal Studies, Cochin University of Science and Technology, Kochi, Kerala

ABSTRACT

"With great power comes great responsibility."

Artificial intelligence is the power bank of the 21st Century. AI has made significant progress in simplifying complex human life. From Chatbots to AI agents, which make autonomous decisions based on preferences and inputs, AI has indeed made significant strides in all arenas of life. Even when AI has the brush in hand, humans decide what to paint. AI is trained with a large amount of databases and preferences, which can thus make accurate predictions and decisions. AI model training is the process of feeding curated data to selected algorithms, which helps the system refine its responses to produce accurate results. Successful AI model training starts with quality data that accurately and consistently represents real-world and authentic situations.

AI is prone to making preferences and prejudices in its output when biased and stereotypical data is provided for training AI models. AI uses predetermined biases for women and other sexual minorities, in cases of hiring bots, suggestive content, and statements. The over-representation of men in the design of these technologies could quietly undo decades of advances in gender equality. Screening of resumes, by Artificial intelligence systems, is the most discussed challenge of Gender discrimination, which leads to automatic rejection of female employees and gender minorities. The system may have an inbuilt capacity to sort and eliminate women according to their age, marital status and even the possibility of a near pregnancy. All this proves that the prevailing gender biases in society can be amplified by automated systems and machine learning technologies. To legally regulate such perpetuation of discrimination by such models would indirectly imply a restriction on discriminatory and detrimental practices and preferences put forth into society. This paper is an effort to analyze the depth of gender discrimination by AI Models and the efficiency of regulatory frameworks worldwide to curtail the same.

Keywords: AI Screening, Gender Bias, Amplification, Hiring Models, AI Training, Automated System, AI Laws.

Artificial Intelligence, or AI, has rapidly moved from being a futuristic concept to an everyday reality. It now influences almost every aspect of our lives, from simple online shopping recommendations to complex, life-changing areas such as recruitment, healthcare, and even policy decisions. While AI holds the potential to enhance efficiency and transform society, it also carries the risk of reproducing and magnifying existing inequalities.

The central concern is that AI itself is not inherently biased; rather, it mirrors the prejudices and stereotypes already existing in human society. Such biases may not be expressed explicitly, yet they can prove influential in behaviour. Machines can learn word associations from written texts and these associations mirror those learned by humans, as measured by the Implicit Association Test (IAT). Because the IAT has predictive value in uncovering the association between concepts, such as pleasantness and flowers or unpleasantness and insects. It can also tease out attitudes and beliefs—for example, associations between female names and family or male names and career. [1]

AI systems can reinforce the pre-existing social biases present in the data they are trained on. Algorithmic bias "ensues when the algorithms employed in machine learning models harbour inherent biases, which are mirrored in their outputs." Whatever prejudices are inherent in our societies, AI tends to replicate and amplify them. If left unchecked, AI can reinforce discrimination and institutionalise social injustice.[2] Feminist theory presents crucial insights and criticisms in addressing algorithmic bias. Particularly, intersectional feminism offers a framework for comprehending how various forms of oppression, including gender, race, class, and sexuality, intersect and compound within algorithmic systems. Feminist data science can challenge conventional methods of data collection, analysis, and interpretation. It can advocate for increased transparency, accountability, and diversity. By adopting feminist principles, feminist data scientists can foster a more just and inclusive environment. [3]

The hiring process is a critical gateway to economic opportunity, determining who can access consistent work to support themselves and their families. LLM (Large Language Model) technologies are likely to make their way into hiring contexts like applicant screening, where

[1] Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 *Science* 183 (2017).

[2] *AI and Its Algorithm Bias and Ethical Implications*, J. Knowledge Learning & Sci. Tech., at 133 (2023), https://www.researchgate.net/publication/397001357_AI_and_its_Algorithm_Bias_and_Ethical_Implications

[3] C. El Morr, *The Need for a Feminist Approach to Artificial Intelligence*, 4 Proceedings of the AAAI Symposium Series 1, 332–33 (2024), https://doi.org/10.1609/aaaiss.v4i1.31812.

intelligent systems are already widely used. Hiring is rarely a single decision point, but rather a cumulative series of small decisions. Predictive technologies can play very different roles throughout the hiring funnel, from determining who sees job advertisements to estimating an applicant's performance to forecasting a candidate's salary requirements. While new hiring tools rarely make affirmative hiring decisions, they often automate rejections. Candidates are deemed by a predictive system not to meet the minimum or desired qualifications needed to move further in the application process. Predictions based on past hiring decisions and evaluations can both reveal and reproduce patterns of inequity at all stages of the hiring process, even when tools explicitly ignore race, gender, age, and other protected attributes. Bias against people of color, gender , and other underrepresented groups has long plagued hiring. Hiring tools that assess, score, and rank jobseekers can overstate marginal or unimportant distinctions between similarly qualified candidates.[4]

The issue of Silicon Ceiling persist in AIhiring methods as an invisible barrier is created by algorithmic systems that affects economic oppurtunities. Unlike the traditional glass ceiling, which involves identifiable and distinct discrimination, the Silicon Ceiling operates even before any formal assessment . One cannot predict the outcome of the algorithmic system, which may caused structural and invisible harm.

A well-documented case reported by Reuters was Amazon’s AI recruitment tool, which analysed resumes by learning from data of previous “successful” candidates. Most of those top-performing resumes were from men, which taught the AI system that men were the better candidates. As a result, CVs from female candidates were rejected, denying them opportunities at the very threshold. The system, which was intended to be neutral, instead perpetuated gender discrimination in recruitment. In a recent study by the University of Pennsylvania and Temple University, USA on LLM use for resume screening in the Netherlands, Lippens showed that, compared to Dutch candidates, GPT was significantly less likely to interview resumes with non-Dutch (including Arab, Asian, Black or White American, and other) names. They found that GPT tended to assign higher scores to men than women, while for the Hiring prompt, the observed trend aligned with gender stereotypes, with women receiving higher scores in majority-woman fields, but not in majority-men ones. It was observed that similar stereotypical trends in terms of race; GPT scored White names higher in occupations where White people

[4] MIRANDA BOGEN & AARON RIEKE, *HELP WANTED: AN EXAMINATION OF HIRING ALGORITHMS, EQUITY, AND BIAS* (Upturn Dec. 2018).

are over-represented.[5]

Similarly, automated facial recognition technology poses challenges by reducing identity to rigid binaries of male and female, leaving no space for non-binary individuals whose identities are erased by algorithms trained on narrow datasets. In everyday contexts, even tools such as Google Translate reflect stereotypes. The way these systems interpret and categorize facial features can be heavily influenced by the cultural biases inherent in their programming and data sets. This can result in a technology that, albeit inadvertently, reinforces stereotypical or culturally insensitive portrayals of certain groups. A stark example of this issue was the 2015 incident where Google Photos erroneously tagged two black individuals as gorillas, highlighting the severe consequences of these biases. Further research has deepened our understanding of these issues, indicating that commercial gender classification algorithms' performances can be significantly influenced by skin color, often demonstrating superior performance for lighter-skinned males and remarkably inferior performance for darker-skinned females. More recently, in 2019, a study probing face detection rates using an array of models identified a substantial bias towards particular ethnicities.

The Automatic Gender Recognition (AGR) systems employ physical markers such as the structure of lips, eyes, and cheeks for gender predictions. AGR predominantly approaches gender classification through a binary lens, categorizing individuals as either male or female. This oversimplified perspective, lamentably, overlooks non-binary or fluid gender identities, inadvertently perpetuating marginalization and eliciting profound societal implications, especially for people who identify as transgender and/or non-binary. Gender identity, as a result of a complex interplay of factors, evolves through continuous social embodiments and representations.[6]

In Natural Language Processing (NLP) and Machine Translation(MT) systems, inherit bias can be seen to get amplified in a large degree. For instance, there is a bias towards males if the conditional probability of "doctor" after "he" is higher than the conditional probability of "doctor" after "she." This bias can occur simply because there are more male doctors in the

[5] Lena Armstrong et al., *The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring*, in Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '24) (San Luis Potosí, Mex., Oct. 29–31, 2024), ACM (2024), https://doi.org/10.1145/3689904.3694699.

[6] Elena Beretta, Cristina Voto & Elena Rozera, *Decoding Faces: Misalignments of Gender Identification in Automated Systems*, 19 J. Resp. Tech. 100089 (2024).

training texts.[7] "When translating from Turkish, which uses gender-neutral pronouns, Google Translate attributes he to the doctor and she to the nurse. '"Positive words like "entrepreneur" or "president" were linked to men, while terms like "nanny' or "lazy" were associated with women. These examples illustrate how machine learning systems internalise human stereotypes and reproduce them, embedding inequality in invisible ways.

The problem is far from isolated. A study by the Berkeley Haas Centre for Equity, Gender, and Leadership, analysed 133 AI systems across industries and found that forty-four per cent exhibited gender bias, while twenty-five per cent reflected both gender and racial bias. This shows that AI is not merely automating processes but also automating judgments, often in a manner that entrenches inequalities.

To address this, we must ask where bias in AI comes from. Bias enters at multiple stages of the machine learning pipeline.

1. Data collection bias arises when training data is incomplete, unrepresentative, or itself drawn from biased sources.

2. Algorithmic bias occurs when the design of the model embeds assumptions that skew outcomes.

3. User interaction bias appears when human inputs reinforce discrimination over time.

There is no one generic approach to fairness, only alternative interpretations, which have implications for mitigating bias. We cannot expect machines to reconcile these differences when society has not, and there will be trade-offs in any chosen approach. Attempts to mitigate algorithmic bias must therefore carefully consider what a fair outcome in any given context should look like and develop strategies accordingly.

How algorithmic bias is likely to be expressed, and the consequences for individuals and groups, is highly context-specific. In some areas, there may be severely detrimental consequences for relatively small numbers of individuals, while in others, there may be relatively minor consequences that are distributed across large subsets of society. The way the AI systems work is sometimes opaque for both technical and proprietary reasons, making

[7] Anna Farkas & Renáta Németh, *How to Measure Gender Bias in Machine Translation: Real-World Oriented Machine Translators, Multiple Reference Points*, 5 Soc. Sci. & Humanities Open 100239 (2022).

scrutiny of bias more difficult. Limited access to the data and systems used by organisations, as well as the use of machine learning algorithms that produce opaque models, impede public scrutiny and the detection of bias.[8]These flaws, individually and collectively, ensure that bias is not accidental but a predictable consequence of technological design rooted in unequal social realities.

The structural context of the AI industry itself compounds the problem. The field is heavily male-dominated. According to the Global Gender Gap Report 2023, only thirty per cent of professionals in AI are women. This imbalance influences research priorities and outcomes. AI tools, for example, often fail to detect or properly diagnose female medical symptoms simply because women's data is underrepresented in medical datasets. Gender minorities also remain largely invisible to mainstream AI development. There has also been little research into how effective existing mitigation techniques are in real-world contexts to remove direct or indirect prejudices.

Given these challenges, fairness, ethics, and accountability must be the guiding principles of AI development. AI should be designed and deployed in a manner that is impartial, unbiased, and equitable. Regulation, therefore, is not just about controlling technology but about safeguarding human dignity and ensuring social justice. Clear rules, standards, and guidelines are necessary to mitigate risks, establish accountability, and protect rights. Internationally, several frameworks have emerged to meet this challenge. The European Union has been at the forefront with its General Data Protection Regulation(GDPR), which strengthens transparency and individual data rights. The OECD(Organization for Economic Cooperation and Development) has issued principles promoting fairness and responsible AI. Canada has developed a Directive on Automated Decision-Making to regulate ethical AI use in government processes, emphasising transparency and fairness.

In November 2021, UNESCO introduced the *Recommendation on the Ethics of Artificial Intelligence*, marking a significant milestone as the first global normative framework aimed at guiding the ethical development and deployment of AI technologies. This comprehensive instrument was adopted by all 193 UNESCO Member States, reflecting a collective commitment to ensuring that AI serves humanity's best interests. At its core, the

[8] Michael Rovatsos, Brent Mittelstadt & Ansgar Koene, *Landscape Summary: Bias in Algorithmic Decision-Making* 3 (Centre for Data Ethics & Innovation 2019).

Recommendation emphasizes the protection of human rights and dignity, advocating for principles such as transparency, accountability, and fairness in AI systems. It underscores the necessity of human oversight, ensuring that AI systems do not undermine human agency or decision-making processes. It stresses that AI systems must be designed and tested to eliminate bias, with fairness audits recommended to systematically check outcomes.

These initiatives demonstrate the urgency of AI regulation. Yet the question remains: are they efficient enough to address the problem? The answer is mixed. While frameworks such as the GDPR, the EU AI Act, and UNESCO's ethical guidelines represent important progress, their efficiency is limited by scope, fragmentation, and weak enforcement. Many of these measures are region-specific, meaning global governance is uneven. Laws often lag behind the pace of technological innovation, leaving gaps where discrimination can thrive. Even where principles are well stated, mechanisms for implementation and enforcement are often underdeveloped. Fairness audits, accountability reviews, and transparency mechanisms are recommended, but not uniformly mandated. This leads to inconsistent compliance and weakens public trust.

In conclusion, Artificial Intelligence is not a neutral or purely technical tool; it is a mirror of our societies, reflecting both progress and prejudice. The path forward requires more than just better data and improved algorithms. It requires building legal and ethical frameworks that are harmonised, enforceable, and proactive. Fairness, accountability, and transparency must not remain aspirational buzzwords but must translate into operational standards with binding effect. Only then can AI evolve as a tool for justice and equality rather than a mechanism for magnifying prejudice. The future of AI governance is not just a question of technology but a question of human values, rights, and dignity.