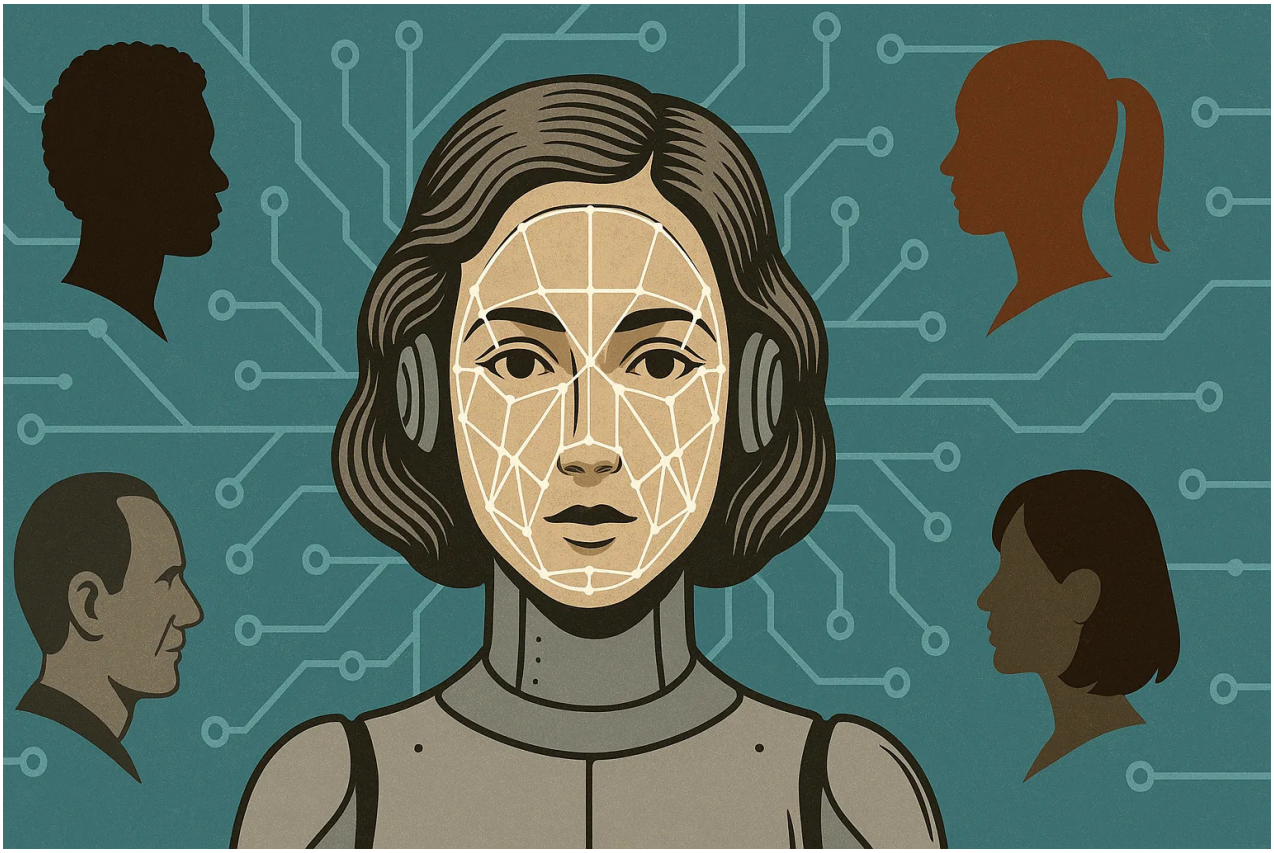
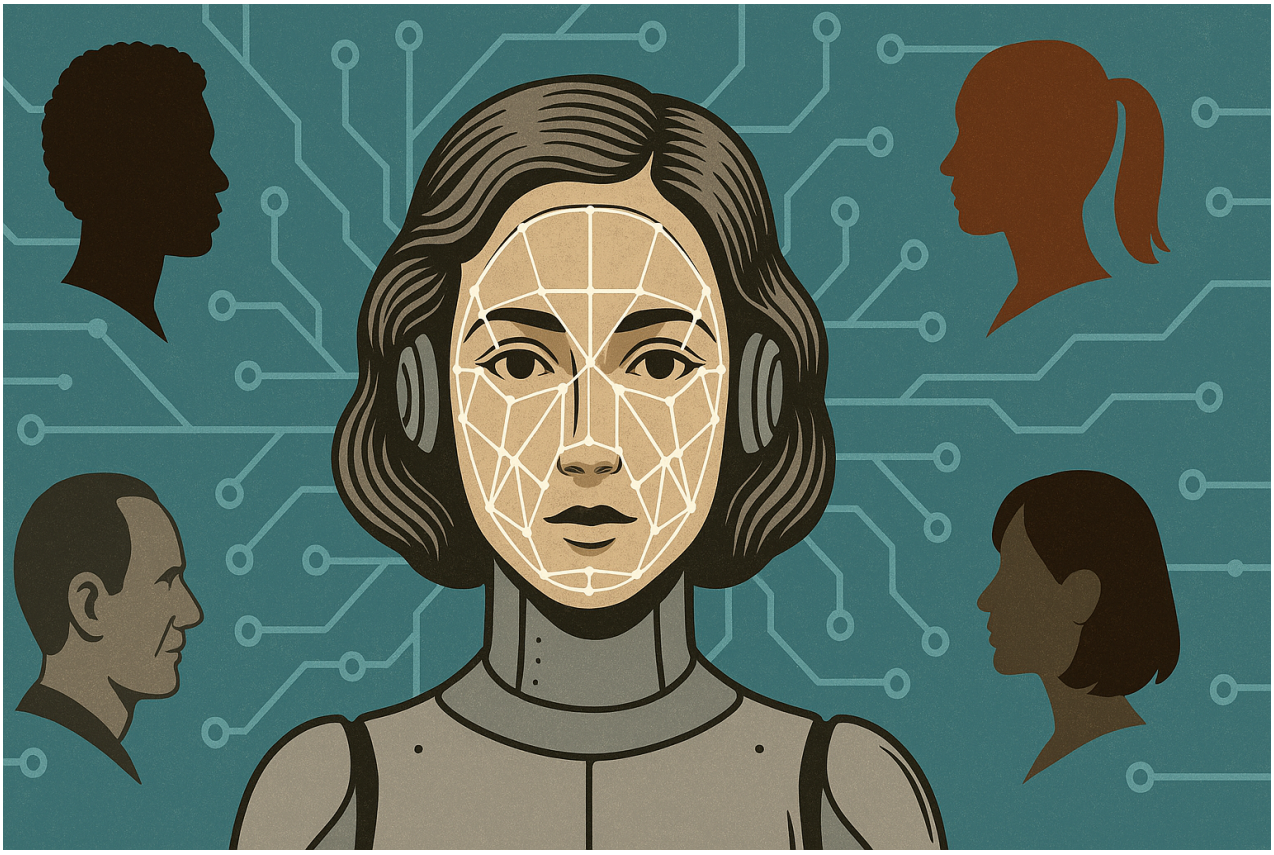
Toward Algorithmic Justice: How Can We Tackle AI Bias?





Artificial intelligence applications often serve as a mirror reflecting the values and culture in which they were developed including both aspirations and, at times, unintended biases.

Despite the widespread portrayal of machines as neutral, algorithms are never divorced from the backgrounds of their developers or the social contexts that shaped them. Reducing AI bias, therefore, requires a deep understanding by tech companies of the societies that use these tools, and of the subtle ways AI can shape individual behavior and decision-making.

While significant efforts are dedicated to improving technical performance and enhancing AI efficiency, ethical considerations are often relegated to the margins. Yet ensuring fairness and transparency demands a comprehensive approach one that starts with model design and continues through post-deployment monitoring.

This approach must include rooted solutions that reflect local specificities, clear disclosures about the nature, sources, and potential biases in training data, and proactive accountability measures.

However, this idealistic vision often runs up against the practical realities of a fast-paced, cost-driven tech industry where ethics may be viewed as a burden unless reinforced by robust regulatory frameworks.

As such, the creation of clear, enforceable legislation remains the most crucial safeguard to ensure that artificial intelligence does not become a sophisticated engine for reproducing gender bias in more subtle and complex forms.

The Reproduction of Biases

In 2014, Amazon began developing an AI-powered recruitment tool to automate résumé screening and identify the best candidates for technical roles. Inspired by its product rating system (from one to five stars), the tool was meant to highlight the most qualified applications.

However, developers soon discovered that the system exhibited a bias against women. It favored résumés that used “masculine” language patterns and downgraded graduates from women’s colleges.

According to a Reuters report, the bias emerged because the algorithm had been trained on ten years’ worth of hiring data data dominated by male applicants effectively teaching the system that men were the “ideal” candidates for tech jobs.

Despite Amazon’s attempts to tweak the algorithm and eliminate these biases, concerns remained about the emergence of new forms of discrimination. In 2018, the company ultimately scrapped the project.

Amazon’s case is not unique. Academic research has revealed similar biases in Midjourney, a generative AI tool for image creation. A study analyzing more than 100 generated images over six months found multiple types of bias.

Among the findings was a racial bias: image prompts for terms like “journalist” or “reporter” consistently produced white-skinned figures, reflecting a lack of racial diversity in the training data. There was also an urban-centric bias, as characters were consistently shown in skyscraper-dense cityscapes, even when no geographic location was specified marginalizing rural representation or alternative settings.

Moreover, the study revealed significant age and gender biases. Younger individuals were typically depicted in lower-skilled roles, while older characters exclusively men were shown in highly specialized positions. Women, by contrast, were consistently portrayed as youthful and flawless, with no visible wrinkles, while the model was more forgiving of signs of aging in men.

Drawing on previous research, UNESCO has warned that large language models (LLMs) and generative AI systems tend to perpetuate gender stereotypes and exhibit what it called “regressive tendencies.” For example, an AI model might generate a story in which doctors are always male and nurses always female reinforcing, rather than challenging, outdated gender norms.

Additional studies have found racial and gender bias in tools like DALL·E 2, which linked high-ranking executive roles to white men 97% of the time.

What About the Arab World?

In the Arab region, AI bias is often treated as a non-issue not because local models are free from bias, but because awareness of the problem remains limited or marginalized. Unlike in the West, where numerous documented cases of algorithmic discrimination against women have emerged, similar scrutiny is largely absent in Arab contexts.

This gap stems in part from a lack of tools capable of monitoring and analyzing how AI models operate within societies. There are few field studies or investigative reports examining AI's impact on key sectors such as employment, healthcare, and media. To date, there are insufficient research platforms or journalistic initiatives focused on this issue, and there's a notable lack of open data to track decision-making processes in either public or private sectors.

The problem is further compounded by the absence of the necessary "knowledge and technical infrastructure." Most AI tools used in the Arab world are developed abroad and deployed locally without adaptation or proper evaluation of the social and cultural contexts in which they operate. This means that potential biases are neither anticipated nor monitored, and AI systems are often treated as "black boxes" opaque and unquestioned.

Another critical shortfall is the lack of metadata or "data about the data" that would allow for scrutiny of the training datasets. Do they represent women? Include diverse geographic regions? Reflect a balance of ages, languages, and cultural backgrounds?

The most honest answer: we don't know perhaps because we haven't even begun to ask.

The real question, then, is not whether digital bias exists in the Arab world, but why we haven't started identifying, analyzing, and addressing it.

The absence of evidence should not be mistaken for technological "innocence." Rather, it highlights the lack of critical tools and transparency needed before intelligent systems are trusted with shaping public policy or guiding vital decisions.

There can be no talk of "just" AI without a conscious, critical environment capable of deciphering algorithmic behavior, exposing embedded biases, and demanding the integration of gender equity as a core requirement in AI design and deployment.

Paths to Confronting Bias

Addressing bias in AI systems starts with acknowledging its existence and understanding its impact on societies. It also requires a clear ethical framework that guides development toward transparency and fairness as advocated by initiatives such as the Algorithm for Equality Manifesto, which stresses the importance of reflecting human diversity in AI models.

Diversifying and improving training data is essential. Datasets must reflect real-world variation in gender, race, age, and background, and follow principles like FAIR to ensure data is Findable, Accessible, Interoperable, and Reusable a key step in minimizing bias.

Regular audits of AI models are another fundamental strategy, as demonstrated by New York City's law requiring audits of automated hiring tools prior to use. Organizations can also leverage open-source tools like AI Fairness 360 to identify and mitigate model bias.

Human oversight plays a pivotal role. Tools such as D-BIAS allow users to identify and alter data relationships that introduce bias, helping create more neutral datasets.

Institutional governance is equally critical. Companies and organizations should establish ethics committees or dedicated monitors and form interdisciplinary teams including legal experts, sociologists, and data scientists to ensure broad, inclusive oversight.

Legislation also plays a vital role. California, for example, is advancing policies aimed at curbing discriminatory practices in AI systems used in employment and healthcare.

Additionally, advanced technical solutions are emerging, such as Affine Concept Editing, which has proven effective in reducing bias gaps in hiring algorithms, along with strategies to refine training data for generative models.

The combination of ethical standards, diverse datasets, continuous auditing, human participation, institutional governance, supportive regulation, and cutting-edge technology offers a solid foundation for building fairer AI systems systems that reflect and serve human diversity equitably.

In the Arab world, however, the lack of open and inclusive datasets underscores an urgent need for regional projects that invest in creating fair, representative data sources, alongside local governance frameworks tailored to the cultural and social specifics of the region.