***Onofriichuk Andrii**, MA student*
***Topachevskyi Serhii**, PhD (Philology), As. Prof.*
Zhytomyr Polytechnic State University

BIAS IN ARTIFICIAL INTELLIGENCE AND ITS INFLUENCE ON SOCIETY

This paper examines academic approaches to analyzing how bias manifests in artificial intelligence (AI) systems and identifies unresolved issues within current research. Numerous AI tools, including large language models, facial recognition systems, and automated decision-making algorithms, tend to reproduce or even amplify existing social biases. For instance, a healthcare algorithm used in the United States was found to systematically underestimate the medical needs of black patients, allocating fewer resources to them compared to white patients with identical health conditions [1, p. 708]. Similarly, commercial facial recognition systems have demonstrated significantly higher error rates for individuals with darker skin tones, particularly women [2, p. 77].

Most researchers explain this phenomenon by pointing to biased training data. AI systems learn from historical texts, images, and databases that already contain societal stereotypes [3, p. 12]. As Bender and her colleagues argue, *"large language models trained on vast internet text corpora inevitably absorb and reproduce the prejudices present in those sources"* [4, p. 611]. Some studies even suggest that AI bias reflects deeper structural inequalities rather than technical flaws alone [5, p. 2].

In this context, different solutions are proposed. One idea is to carefully curate training datasets to remove or balance biased examples [3, p. 18]. Another approach involves algorithmic fairness techniques, such as adjusting model outputs to ensure equal treatment across demographic groups [6, p. 236]. Some researchers also argue that bias mitigation must go beyond technical fixes, as it concerns justice, representation, and human rights [5, p. 45].

Bias in AI is examined through outputs drawn from varied domains — such as hiring platforms, credit scoring systems, healthcare algorithms, and content moderation tools — where automated decisions may implicitly reinforce racial, gender, or socioeconomic inequalities [1, p. 710; 7, p. 4].

But even with all this research, there are still gaps. Most studies focus on Western contexts and English-language systems [5, p. 8]. AI systems deployed in non-English or non-Western settings are rarely discussed, even though they face similar or even more severe bias challenges. Additionally, most researchers examine large commercial systems, such as those from Google, Microsoft, or Amazon. Smaller, local, or open-source models are often overlooked, although they are also used in real-world applications [7, p. 6].

The frequent reproduction of stereotypes in AI output stems not only from statistical patterns in training data but also from entrenched cultural norms. As Safiya Noble observes in her study of algorithmic oppression, *"algorithms are not neutral — they are shaped by human values and institutional priorities"* [8, p. 17].

Another important question is: where does the bias come from? Researchers continue to debate whether bias stems primarily from biased training data, from model architecture, or from the way systems are deployed in unequal social contexts. A comprehensive survey by Mehrabi

et al. identifies multiple sources of bias, including historical bias, representation bias, measurement bias, and evaluation bias [3, p. 14]. Obermeyer et al. demonstrate that even when sensitive attributes like race are explicitly removed from training data, algorithms can learn proxies for those attributes that perpetuate discrimination [1, p. 712]. Fixing this means working on data, model design, and deployment practices simultaneously.

Different types of AI systems also handle bias differently. Traditional rule-based systems are more predictable but less flexible, while deep learning models achieve higher accuracy but are harder to interpret and control [6, p. 240]. Large language models, for example, excel at fluency but remain weak in fairness and safety [4, p. 615]. As Raji and her colleagues note, *"developers should conduct regular audits of their systems to identify and mitigate bias before deployment"* [9, p. 6].

Some researchers go even further, arguing that AI bias extends beyond technical performance. These systems are increasingly used in education, healthcare, criminal justice, and public services. In the criminal justice system, for example, risk assessment algorithms like COMPAS have been shown to overestimate recidivism risk for Black defendants while underestimating it for white defendants [6, p. 238]. Cathy O'Neil warns that many AI systems are *"weapons of math destruction"* that *"punish the poor and marginalized while hiding behind claims of mathematical objectivity"* [10, p. 8]. That is why fairness in AI matters. It is not just a matter of getting predictions right — it reflects a deeper need to respect human dignity and equality.

Users from marginalized communities may experience subtle but repeated harm when encountering biased AI outputs, particularly in contexts where automated decisions shape access to opportunities — including job applications, loan approvals, medical diagnoses, and legal assessments [5, p. 6]. As Ruha Benjamin writes, the use of predictive algorithms can create a *"new Jim Code"* — a form of high-tech discrimination that seems neutral but systematically disadvantages already vulnerable populations [5, p. 50].

This broader view of bias in artificial intelligence has led researchers to consider not only accuracy but also how biased systems affect people's lives. If AI tools keep linking certain groups to negative outcomes, they help spread a quiet but steady form of discrimination. This becomes a real problem when automated systems are used in schools, government, or courts, because they influence how society allocates opportunities and treatment. As Joy Buolamwini and Timnit Gebru argue, *"the failure to ensure equitable performance of facial recognition systems across intersectional groups can lead to misidentification, exclusion, and harm"* [2, p. 89].

To deal with this, some experts suggest making AI systems more transparent and accountable. Raji and her colleagues propose an *"end-to-end framework for internal algorithmic auditing"*, emphasizing that *"accountability cannot be an afterthought; it must be built into the AI lifecycle from the start"* [7, p. 10].

Another idea is to use training data that includes diverse and balanced representation across all social groups. Mehrabi et al. state that *"many AI systems are trained on data that mostly represent dominant groups, especially in science, business, and technology"* [3, p. 22]. If developers use more balanced data, the results can be more fair. However, as Bender and colleagues caution, *"simply balancing datasets is insufficient; researchers must also consider the broader social contexts in which data are generated and interpreted"* [4, p. 620].

This study explores the emergence of various biases in AI systems. It proposes context-aware strategies for bias mitigation, with specific attention to how biased AI outputs affect public-facing communication, social trust, and equal access to opportunities.

Finally, as Ruha Benjamin concludes, *"the question is not whether we will use algorithms to make decisions, but how we will ensure that they do not reproduce the very inequalities we claim to overcome"* [5, p. 58]. Should AI focus on accuracy, efficiency, or fairness? There is no single answer, but it is clear that bias in artificial intelligence is a deep issue that connects technology, ethics, and social justice. More research — especially in non-Western languages and local contexts — is needed to ensure that AI systems treat all people fairly.

Bias asymmetries in AI output reveal systemic imbalances in how intelligent systems encode and circulate social hierarchies. Remediation efforts must extend beyond algorithmic refinement to include sustained attention to fairness, transparency, and inclusive representation in AI development and deployment practices.

REFERENCES

1. Obermeyer Z., Powers B., Vogeli C., Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations // Science. — 2019. — Vol. 366(6464). — P. 708—712.
2. Buolamwini J., Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification // Proceedings of FAT*. — 2018. — P. 77—91.
3. Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A. A survey on bias and fairness in machine learning // ACM Computing Surveys. — 2021. — Vol. 54(6). — P. 1—35.
4. Bender E. M., Gebru T., McMillan-Major A., Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? // Proceedings of FAT*. — 2021. — P. 610—623.
5. Benjamin R. Race after technology: Abolitionist tools for the new Jim code. — Cambridge : Polity Press, 2019. — P. 1—60.
6. Selbst A. D., Boyd D., Friedler S. A., Venkatasubramanian S., Vertesi J. Fairness and abstraction in sociotechnical systems // Proceedings of FAT*. — 2019. — P. 236—247.
7. Raji I. D., Smart A., White R. N., Mitchell M., Gebru T., Hutchinson B., et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing // Proceedings of FAT*. — 2020. — P. 33—44.
8. Noble S. U. Algorithms of oppression: How search engines reinforce racism. — New York : NYU Press, 2018. — P. 15—30.
9. Raji I. D., Gebru T., Mitchell M., Buolamwini J., Lee J., Denton E. Saving face: Investigating the ethical concerns of facial recognition auditing // Proceedings of AIES. — 2020. — P. 145—151.
10. O'Neil C. Weapons of math destruction: How big data increases inequality and threatens democracy. — New York : Crown, 2016. — P. 1—20.