Human TECHNOLOGY

ISSN: 1795-6889

https://ht.csr-pub.eu

Volume 21(2), September 2025, 246–250

SOCIAL IMPACT OF DATA BIAS IN ARTIFICIAL INTELLIGENCE MODELS

Adam Wojciechowski
Lodz University of Technology
Poland
ORCID 0000-0003-3786-7225

Kristiina Korjonen-Kuusipuro
South-Eastern Finland University of Applied Sciences
Finland
ORCID 0000-0002-8528-0237

Abstract: *The growing popularity and scope of application of artificial intelligence models requires attention to be paid to their reliability and the explainability of the reasons behind the results and recommendations they generate. Widespread automation driven by the development of AI may have serious social consequences, and data bias may be the cause of unethical and socially unacceptable tendencies in models. This can manifest itself both at the stage of data collection and processing, but also during model training or implementation. It is the duty of the scientific community not only to pay special attention to exposing the biases of models, but above all to track down unacceptable implementations and strive to take into account the social consequences of the technological solutions being developed. In this study, we will look not only at the causes of bias, but also at examples with significant social implications and measures aimed at eliminating bias and its social consequences.*

Keywords: *data bias, social consequences, bias examples, bias mitigation.*

DOI: https://doi.org/10.14254/1795-6889.2025.21-2.0

INTRODUCTION

The popularity and declared effectiveness of artificial intelligence models have become the driving force behind contemporary technological and social changes. The automation of industrial processes and the gradual reduction of human involvement in decision-making, as well as the increasingly widespread generation of opinion-forming content and advertising, are diverting humanity's vigilance and attention away from the bias of AI models. The classic AI evaluation system, based on measures such as accuracy, effectiveness and precision, does not allow for an assessment of their reliability, especially given the rather limited access to the learning algorithms and data used in AI model training.

It is important to be aware that artificial intelligence models attempt to find patterns in the space of features and accompanying observations. What happens when features or observations are underrepresented or do not take certain perspectives into account? What tendencies does a model have when biased data is used? Intuitively, one could respond that data-dependent artificial intelligence models take on the characteristics of the data itself and its interpretation. Regardless of the quality of the data, the vulnerability of artificial intelligence models may result from the learning process itself, e.g. the attention mechanism or cost function allow the model's output to be biased in order to encode the creator's intentions. Model biases may also result from runtime data and not just the data on which the model was trained.

However, this does not change the fact that data is the most common cause of model bias, which is why we will take a closer look at different types of bias. We can distinguish several types of bias. The most common is selection bias. It occurs when the training data is not representative of the population. Outlier bias is a certain complement to this, where we focus on general trends and ignore rare and unusual events, distorting the spectrum of possible observations. Equally common, although users are much less aware of it, is confirmation bias. It occurs when the AI model's behavior confirms our existing attitudes or beliefs, especially since we usually like to be right, and algorithms have learned to confirm this. It is often similar to historical bias, where data is overly dependent on historical prejudices or stereotypes. Measurement bias, on the other hand, occurs when the data collected systematically fails to reflect the actual variables and dependencies of the process. Either we focus on selected (most frequent and therefore accessible – availability bias) aspects, ignoring those we consider less important, or we do not consider the full spectrum of variants, focusing only on successes and not analyzing cases of failure (survivorship bias).

Despite numerous deficits in the available data sets, stakeholders seem so strongly guided by their own priorities that data bias is usually not overestimated. On the one hand, scientists are competing, sometimes blindly, to achieve the best results, but without paying due attention to the interpretability/explainability of the solutions obtained. Overfitting the model does yield better results, but only on the data set used. Verifying the solution on a different data set usually reveals its imperfections. However, examining the origin of the result requires reflection and time, which not everyone is willing to invest. Driven by the desire for profit, industry representatives are trying to commercialize the available knowledge without concern for the ethical aspects of AI models. It is not difficult to imagine a situation where a potentially greater material benefit would encourage the creators or users of AI models to manipulate the data or the model itself in order to achieve their intended goal. On

the other hand, ordinary users view everything with a mixture of fear of change and possible job loss, but on the other hand, they want to opportunistically use the available tools to make their work easier. However, they tend to forget that they are acting as unpaid laborers, sacrificing their free time to feed the algorithms of large language models or social media.

Let us therefore look at selected examples of bias in the available solutions and then consider whether we can influence this. As researchers or journal editors, can we increase our diligence in striving to reduce the bias of AI models and increase their reliability?

DATA BIAS EXAMPLES

One area that is not free from critical examples of AI bias is healthcare. Rickman et al. (2025) analysed the tendency of selected LLMs to automatically summarize long-term care cases in a sample of the British population. The performance of four language models (Llama3, Gemma, T5, BART) was verified among more than 600 elderly people. Gender bias was observed in some models, with summaries of men focusing more on physical and mental issues than those of women and being formulated in a more direct manner. In this way, women's needs may be overlooked more often than men's. Unfortunately, examples of gender-based social injustice are much more common. According to The Times[1], Apple credit card holders (formally managed by Goldman Sachs) have repeatedly reported that rating algorithms granted men credit limits many times higher than women, even when women could demonstrate higher incomes and greater financial assets. Even Apple co-founder Steve Wozniak admitted to this, having been given a 10 times higher limit than his wife, despite their full community of property. Dastin (2022) demonstrated that companies such as Amazon have become victims of their own success. The automation mechanism, built on AI algorithms, which was the foundation of Amazon's market position, was applied to recruitment. Unfortunately, the algorithm favored men and depreciated applications submitted by women. Data bias, manifested in the reproduction of gender stereotypes, is present in popular graphics tools. Popular generative image creation tools, such as DALL-E and Stable Diffusion, when asked to generate images of senior managers (e.g., CEOs), overwhelmingly use images of white men. At the same time, requests to generate images of nurses or cleaners are depicted using women or minorities. Yu et al. (2022) exposed an algorithmic fairness metric that was used by LinkedIn for recruitment recommendations. The algorithm favored men over women with the same qualifications. Fortunately, LinkedIn improved its algorithm.

Healthcare is a key social area where manifestations of racism can be found. Obermeyer et al. (2019) exposed the racist tendencies of the American healthcare system, which uses commercial algorithms to support treatment decisions. The authors showed that the algorithms reduced the chance of black people qualifying for treatment by 50%. The algorithm used the cost of treatment as an indicator of health needs. Lower treatment costs for black people compared to white people meant that the algorithm treated black people as healthier and therefore referred them for additional tests less often. Manifestations of racism can also be found in systems that support the justice system. An example is the COMPAS

[1] https://time.com/5724098/new-york-investigating-goldman-sachs-apple-card/

algorithm, now operating under the name Equivant, created by Northpointe to predict the risk of crime in American courts. A 2016 report by ProPublica[2] found that black defendants were incorrectly assessed as high risk (45%) compared to white defendants (22%). On the other hand, white defendants were significantly underestimated as low risk, despite reoffending. In turn, Buolamwini et al. (2018) published a paper exposing critical data bias in commercial facial recognition systems used by companies such as IBM and Microsoft. The authors showed that the classification error for black women was 34.7%, while for white women it was only 0.8%. Seemingly innocent recommendation environments are also not free from manifestations of racism. A recently published platform for testing hairstyles[3], which was supposed to assess a person's intelligence and professionalism on this basis, rated photos of black women worse. Summarizing the examples documented in literature and reports, one can only imagine how many manifestations of bias in AI data and models can be expected.

CONCLUSIONS AND IMPLICATIONS

So, the question arises: how can bias be eliminated or prevented? The most obvious way to prevent bias is to ensure diverse and representative data in the model learning process, both in terms of the problem perspective and demographics. Reviewers and editorial committees of journals have a special role to play in this field. Attention to research methodology, appropriate selection and diversification of data sets are among the basic measures that are the responsibility of the scientific community. Being aware of the lack of explainability of contemporary artificial intelligence models, work on broadly understood explainable AI should be intensified. Regardless of this, we can expect a proper discussion and critical reflection on the social implications of the models being developed and the data sets used. We still feel that such critical discussion is insufficient.

However, nothing can replace continuous monitoring to detect emerging biases. In all scenarios that may have ethical or legal implications, it is worth keeping humans in critical decision-making loops. At the same time, for existing AI models, one may not be aware of their biases. In such cases, numerous bias verification tools can be used. Popular tools include: Google's What If-Tool[4], Amazon SageMaker Clarify[5], Fiddler AI[6], IBM AI Fairness 360[7], Microsoft Fairlearn[8] and Credo AI[9]. They use generally accepted fairness metrics, adversarial testing and counterfactual analysis to observe how changes in input data affect AI model predictions. A universal mechanism is model explainability, which should allow for a better understanding of decision-making criteria. With access to decision-making criteria and thresholds, it is always possible to verify how they affect fairness or facilitate informed decision-making.

[2] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[3] https://janicejnice.substack.com/p/i-tested-ai-on-different-black-hairstyles
[4] https://pair-code.github.io/what-if-tool/
[5] https://aws.amazon.com/sagemaker/ai/clarify/
[6] https://www.fiddler.ai/
[7] https://research.ibm.com/blog/ai-fairness-360
[8] https://fairlearn.org/
[9] https://www.credo.ai/

However, nothing can replace continuous monitoring to detect emerging biases. Furthermore, in all scenarios that may have ethical or legal implications, it is worth keeping humans in the loop for critical decisions.

REFERENCES

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296-299). Auerbach Publications.

Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads, (2017) https://www.propublica.org/article/facebook-ads-age-discrimination-targeting

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

Rickman, S. Evaluating gender bias in large language models in long-term care. *BMC Med Inform Decis Mak* **25**, 274 (2025). https://doi.org/10.1186/s12911-025-03118-0

Yu, Y., & Saint-Jacques, G. (2022). Choosing an algorithmic fairness metric for an online marketplace: Detecting and quantifying algorithmic bias on LinkedIn. *arXiv preprint arXiv:2202.07300*.

Authors' Note

All correspondence should be addressed to:
Adam Wojciechowski
Lodz University of Technology
Institute of Information Technology
al. Politechniki 8, 93-590 Łódź
adam.wojciechowski@p.lodz.pl

Human Technology
ISSN 1795-6889
https://ht.csr-pub.eu