Bias in artificial intelligence

SELECTED RESEARCH FINDINGS ON STEREOTYPES IN ARTIFICIAL INTELLIGENCE

Sélène Coutant, Alessia Cortina

The article presents a selection of current studies on the reproduction of stereotypes by generative AIs. The main topics are gender, race, living with disabilities, sexual orientation, and age.

Artificial intelligence (AI) is increasingly being used to provide information, explain concepts, design images and sounds, and perform a wide range of tasks for people. However, its design and application bring particular challenges. One of these is the reproduction of stereotypes, as the algorithms behind generative AIs often exhibit bias. This bias can occur at every stage of their development, from design and modelling decisions to data collection and processing and the context of use (UNESCO, 2024). The scale, adaptability and complexity of AI development pose significant challenges in reducing stereotypes and preventing harm to individuals and society. This article summarises a selection of recent studies on the most common text-to-text (TTT) and text-to-image (TTI) AI models, highlights the most frequently investigated stereotypes, and provides an overview of their impact on inclusion and diversity.

GENDER BIAS

Studies on bias through AI often focus on disparities in the representation of men and women. Women are significantly underrepresented in images generated by AI models such as *Midjourney*, *Stable Diffusion*, and *DALL-E 2*, as an analysis of the 3 AI applications shows (*Midjourney*: 23% women vs. 77% men, *Stable Diffusion*: 35% women vs. 65% men, *DALL-E 2*: 42% women vs. 58% men; Zhou et al., 2024).

Women in the workplace

A study by the University of British Columbia examined 8,000 images generated by the AI models *Midjourney*, *Stable Diffusion* and *DALL-E 2* based on English text prompts representing 1,016 occupations (Zhou et al., 2024). It found that men appeared more frequently than women in all occupational groups examined. Zhou and her colleagues criticise that such bias can have a negative impact on the professional development, especially that of young women. This tendency decreases somewhat for professions that require longer training.

Luccioni and her team confirm the stereotyping of "female occupations". They tasked the AI *DALL-E 2* with creating a "photo portrait of a (x) (y) at work", where x stood for ethnicity and y for gender. The occupations depicted were compared with a list of 146 occupations from the U.S. Bureau of Labor Statistics. The analysis revealed that women were, on average, 27% less frequently

(UNESCO, 2024, p. 13)

Ill. 1: AI-generated open-ended stories show clear differences between the Global South and North in terms of subjects

represented in occupations such as clerk, data entry keyer, and real estate broker. Conversely, women appeared more frequently in occupations like singer (36%), cleaner (20%), and dispatcher (19%) (Luccioni et al., 2023). This disparity is particularly evident in occupational fields characterised by cultural and gender-specific stereotypes. For example, according to the aforementioned UNESCO study, British men were depicted in diverse occupations such as driver, caregiver, bank clerk and teacher, while British women were more frequently depicted in occupations such as prostitute, model and waitress, which points to misogynistic tendencies in the depictions. Similarly, according to the study, Zulu men appeared as gardeners, security guards and teachers, whereas Zulu women were predominantly depicted in domestic or service-oriented roles such as domestic servant, cook or housekeeper (UNESCO, 2024).

Visual representation

In the previously cited study by Zhou et al., in which images were generated with *Midjourney*, *Stable Diffusion* and *DALL-E 2*, men were often depicted as older and more experienced than women, which, according to the authors, puts them in an alleged position of authority. Women, by contrast, were more likely to smile and show more positive emotions, while men were often portrayed as neutral or angry. This gender-specific representation and pattern was found across all 3 AI models analysed (Zhou et al., 2024).

UNESCO (2024). Challenging systematic prejudices: an investigation into gender bias in large language models. *Available at: https://unesdoc.unesco.org/ark:/48223/pf0000388971 [8.10.24]*

Zhou, Mi, Abhishek, Vibhanshu, Derdenger, Timothy et al. (2024). Bias in generative AI. *ArXiv abs/2403.02726.*

Luccioni, Alexandra Sasha, Akiki, Christopher, Mitchell, Margaret et al. (2023). Stable bias: analyzing societal representations in diffusion models. *ArXiv abs/2303.11408.*

RACIAL BIAS

Researchers at New York University Abu Dhabi were able to demonstrate a clear bias against People of Colour (PoC) and their underrepresentation in generative AI depictions. Giving the image-generating AI *Stable Diffusion* a racially neutral prompt, just under half of the 10,000 portraits generated showed White people, a further third showed Black people and only 8% of the images showed people perceived as Asian or Indian (AlDahoul et al., 2024). Zhou et al. from the University of British Columbia in Canada found similar trends, with Black people less frequently depicted across the 3 TTI-programs *Midjourney*, *Stable Diffusion* and *DALL-E 2*. Interestingly, *Stable Diffusion* generated over 50% of its images representing Asian people (Zhou et al., 2024).

Computer scientists at NYU Abu Dhabi investigated the homogenisation of marginalised groups by AI. They generated 10,000 images per race using the AI model *Stable Diffusion* and compared the resulting images with those produced by a diversity-trained algorithm. The portraits generated with the latter showed significantly more differences and variation in the representation of the faces; especially for people from the Middle East and Latin America (AlDahoul et al., 2024). Generalisation can not only be observed on a visual level, but also in written text. For a UNESCO study, 4,000 open-ended stories generated with the TTT-AI *Llama 2* were analysed. The study revealed clear differences between the Global South and North. While the stories set in the Global South mainly contain themes such as community, family, and village, and focus particularly on hardship, labour, education, and dreams, stories set in the Global North are more lighthearted and address themes such as love, feelings, and exploration (Ill. 1; UNESCO, 2024).

Professional stereotypes

A disregard of PoC in the results of generative AIs can also be observed in an occupational context, as a study by NYU Abu Dhabi shows. AlDahoul et al. used *Stable Diffusion* to generate images of 25 different professions, finding

Ill. 2: Studies show: AI-generated images of politicians show mostly White men

Ill. 3: In the context of well-paid professions, TTI-AIs primarily generate images of men with light skin tones, here: CEOs

that in 18 of the 25 professions, White people were depicted more frequently than all other groups combined. In addition, some occupations where White people appeared less often were typically low-paying jobs (AlDahoul et al., 2024).

These racially motivated prejudices have also been proven in other studies. The journalists Nicoletti and Bass, for instance, analysed 300 AI-generated images of 7 professions that are typically considered low-paying jobs in the USA and 7 professions that are considered well-paid. They found that men with light skin tones appear in the majority of the images in all well-paid occupations (e.g., politician, lawyer, judge and CEO; Ill. 2 and 3[1]), whereas people with dark skin tones were represented in low-paying occupations such as fast-food worker and social worker. The authors state that the images do not reflect reality, as a comparison with actual data from the U.S. Bureau of Labor Statistics revealed. People with darker skin tones were depicted in 70% of the generated images of fast-food workers, although real-world occupational data state that 70% of fast-food workers in the USA are White. The result is a stereotypical representation of reality in which inequalities appear more pronounced than they actually are (Nicoletti & Bass, 2023).

Bias through linguistic associations

AI-generated content can reinforce existing social associations between words and images, promoting racist stereotypes, as a study by NYU Abu Dhabi has shown. The researchers had the AI *Stable Diffusion* create portraits of people with the following attributes: winner, beautiful, intelligent, parent, sibling, terrorist, poor, criminal. The results show that, depending on the attribute, the AI-generated images of people with different skin tones and ethnicities: attributes related to success and attractiveness (winner, beautiful and intelligent) and family-related descriptions (parent, sibling), predominantly generated White people. In contrast, the prompt "terrorist" mainly depicted Middle Easterners, but no White people. In addition, the attributes "poor" and "criminal" were mainly linked with Black people (Al-Dahoul et al., 2024). Nicoletti and Bass came to similar conclusions investigating images created with the prompts "inmate" and "terrorist". Of the more than 5,000 images of inmates generated with *Stable Diffusion*, more than 80% showed PoC. The prompt "terrorist" resulted exclusively in portraits of men with dark facial hair and often with headgear, which clearly correspond to Muslim stereotypes. The authors compared these figures with the actual statistics for the USA, which show that less than half of people in prison in the USA are Black – i.e. significantly fewer than the AI claims (Nicoletti & Bass, 2023). It can therefore be concluded that racial stereotyping in AI-generated content does not reflect the actual data and distorts reality.

AlDahoul, Nouar, Rahwan, Talal & Zaki, Yasir (2024). AI-generated faces influence gender stereotypes and racial homogenization. *ArXiv abs/2402.01002.*

Nicoletti, Leonardo & Bass, Dina (09.06.2023). Humans are biased. Generative AI is even worse. *Stable Diffusion's* text-to-image model amplifies stereotypes about race and gender – here's why that matters. *Bloomberg Technology + Equality. Available at: https://www.bloomberg.com/graphics/2023-generative-ai-bias/?srnd=undefined [20.01.25]*

REPRESENTATION OF PEOPLE WITH DISABILITIES

Researchers at the Association for Computational Linguistics used the *BERT* (Bidirectional Encoder Representations from Transformers) language representation model to investigate the representation of disabilities, in particular visual and hearing impairments. The *BERT* model aims to simplify texts to improve their comprehension for users. The study tested the language model with cloze

text tasks. The findings indicate that people with disabilities were often portrayed negatively, frequently associated with illness and that they were less frequently described as friendly or competent (Herold et al., 2022).

Herold, Brienna, Waller, James & Kushalnagar, Raja (2022). Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In Sarah Ebling, Emily Prud'hommeaux & Preethi Vaidyanathan (Eds.), Ninth workshop on speech and language processing for assistive technologies (SLPAT-2022) (p. 58-65). *Dublin: Association for Computational Linguistics.*

REPRESENTATION OF SEXUAL ORIENTATION

The aforementioned UNESCO study found that AI models that work without human feedback produce homophobic content. This is particularly the case when the models are asked to complete sentences about a person's gender or sexual identity. Negative content about homosexual people was generated in about 70% of cases with *Llama 2* and in about 60% of cases with *GPT-2*. For example, *GPT-2* generated sentences such as "Gay people were regarded as abnormal, and it was believed that their lives would be difficult to live." The content generated by *Llama 2* contained sentences such as "Gay people were regarded as the lowest in the social hierarchy" (UNESCO, 2024, p. 10). *ChatGPT*, by contrast, generated positive or neutral content for all subjects in over 80% of cases. This shows that large language models fine-tuned with human feedback generate less negative bias for subjects outside heteronormativity than AI models working without human feedback, even though they are not completely bias-free (UNESCO, 2024).

AGE REPRESENTATION

One group of people that is often ignored when considering ethical issues in connection with AI are senior citizens. Age discrimination can occur due to bias in the data basis but also at the developer or user level (Stypinska, 2023). There are no current studies available that demonstrate age bias. However, in their review paper, Chu and her team were able to identify possible discrimination based on age, also known as "digital ageism", primarily due to inadequate representation and intersectional inequalities (Chu et al., 2023). The authors of both studies encourage further research in this area to address potential discrimination.
Explicit studies on the representation of children and young people and their possible stereotyping through AI are not yet available. However, media pedagogical approaches can be found that deal specifically with bias in AI in children and adolescents (e.g., Vartiainen et al., 2024).

Stypinska, Justyna (2023). AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. AI & Society, 38, *665-677.*

Chu, Charlene, Donato-Woodger, Simon, Khan, Shehroz et al. (2023). Age-related bias and artificial intelligence: a scoping review. Humanities and Social Sciences Communications, 10, *510.*

Vartiainen, Henriikka, Kahila, Juho, Tedre, Matti et al. (2024). Enhancing children's understanding of algorithmic biases in and with text-to-image generative AI. *New Media & Society.*

CONCLUSION

The research makes it clear that even subtle bias in AI-generated content – be it in terms of gender, race, disability, or sexual orientation – can amplify stereotypes and influence public perception. The underrepresentation of non-White, non-male and disabled people is a common issue across all the studies described. The reproduction of stereotypes by algorithms can have far-reaching consequences from reinforcing the consolidation of discriminatory structures to affecting the labour market, for example, if AI is used in the selection of candidates (UNESCO, 2024), or in healthcare if AI is used for diagnostic purposes (e.g., Seyyed-Kalantari et al., 2021). However, the current studies also suggest that regular monitoring and incorporating independent human feedback into AI models can make a difference. For example, the latest version of *ChatGPT*, optimised with human feedback, is less likely to reproduce stereotypes than AIs without human feedback (UNESCO, 2024). Since AI reproduces biases from data sets, it is important to sensitise people to existing and even historically rooted biases and to consciously use diverse and inclusive data. Furthermore, there are examples of programs that uncover bias in algorithms (e.g., *BiasAsker* in Wan et al., 2023), or experiments with models trained on inclusivity or diversity that show good results (e.g., in AlDahoul et al., 2024).

Seyyed-Kalantari, Laleh, Zhang, Haoran, McDermott, Matthew et al. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, *27, 2176-2182.*

Wan, Yuxuan, Wang, Wenxuan, He, Pinjia et al. (2023). BiasAsker: Measuring the bias in conversational AI system. *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2023). New York: Association for Computing Machinery, 515-527.*

NOTE

[1] *The pictures were generated using* Midjourney *(Ill. 2) and* Stable Diffusion *(Ill. 3) with the prompt "Group of politicians/CEOs".*

THE AUTHORS

Sélène Coutant, M. A. Cultural and Cognitive Linguistics, and Alessia Cortina, B. A. English and Political Science, are freelancers at the IZI, Munich.