NeuroConverse Journal, Issue 1, March 2024

Unveiling Neuronormative Bias: Exposing an Unconscious Bias in Generative AI and Their Impact on Neurodivergent People

Claudia Lemke, Martin Bloomfield, Florian N. Herfurth

Algorithms increasingly dominate our world – it is the **algorithmisation of our lives**. In our day-to-day lives, we use digital technologies, often now enriched with artificial intelligence (AI) algorithms that are supposed to give us the promise of optimising our lives. Even if the idea of a computerised intelligence that can simulate our human cognitive capabilities is a noble goal that the great British mathematician and computer scientist Alan Turing (Turing, 1950) envisaged, we see the current AI-based algorithms more as a mirror of our imperfect world, laden with prejudice, injustice, inequality, and unfairness.

The ability to learn through training with the nearly infinite internet data makes AI-based machine learning algorithms and neural networks the brightest stars in the current AI sky. It is not only these algorithms' output that shows several cognitive biases (so-called algorithmic bias) that produce harm and discrimination against marginalised groups; it is also the nature of the training which shows the prevailing injustices and exploitation in our real or natural world, as a glimpse in a mirror. For example, training such systems requires humans to assess data, partly to train the algorithms to learn less cognitive biases. However, this labour force of the global South is being exploited, especially by the Western hemisphere and our competition-driven belief in wealth generation. Not only are they barely paid (Perrigo, 2023), but no one cares about their mental health. These people's work is to watch videos, listen to audio or look at images and texts from the internet to train these algorithms and decide whether data are appropriate. To do this, they look at violent or inhuman images, read hate speech or fake news, and see child abuse, to name but a few of the horrors they must subject themselves to. There is also a third problem with the increasing alogrithmisation of our lives by AI systems that cannot be ignored: the rising energy costs for developing and operating these systems (Luccioni et al., 2023) in massive data centres worldwide.

However, let us go back to the algorithmic bias issue we are exposed to through the many different AI systems in our everyday lives. Biases in AI systems are threefold: Data, as an artificial model of our reality, depicts our cognitive conscious and unconscious biases and limitations inherent in our societies and cultures. AI-based systems that learn by training from real data reinforce these. Additionally, software engineers also plant their view of the world in

the development of algorithms and thus reinforce specific world views, especially if the teams developing AI algorithms are not diverse, either in an ethnic, gender or cultural sense. The third area relates to the fact that the AI systems themselves cannot apply or use common sense when interpreting data (Nishant et al., 2023) that has yet to be processed in the learning process. All three dimensions lead to the already-mentioned effects of reinforcing prejudices against humans.

If we look at a specific category of AI-based systems—**Generative AI**—we are confronted with a sometimes extreme amplification of biases. These general-purpose systems use our human language as input to generate new content, which we then consume, such as text, images, videos, or audio. These models are based mainly on AI algorithms from machine learning, natural language processing, and neuronal networks (Fengchun & Wayne, 2023). According to estimates, the total volume of data and information created, stored, and consumed globally via the internet will probably amount to around 181 Zettabytes by 2025 (Statista, 2024). Tech companies with their dominant market position have invested massively in the construction of data centres to process this data, therefore enabling the training of such transformation models with Internet data, sometimes in deliberate disregard of copyright and data protection of personal data and under exploitative working methods, thus creating the current AI hype around the leading systems that we all use every day today.

Large language models (LLM) such as ChatGPT are among the most prominent representatives of this generative AI. By now, we know that these language models contain biases such as gender, race, or religion (Tamkin et al., 2021; Blodgett et al., 2021; Lucy & Bamman, 2021), which is hardly surprising as these models replicate the prejudices already ingrained within our language. In addition to these biases in generative AI, we now also see other less obvious or more hidden biases, such as the brilliance bias, which addresses the inequalities or distribution between men and women in their representation in the various research disciplines (Shihadeh et al., 2022).

Despite these already-known forms of biases driven by AI-based algorithms, we are observing ongoing discussions about the transformative nature of these systems across many sectors and industries. They show us in a seductively simple way what knowledge creation and knowledge processing of the future should look like. Therefore, all knowledge-based processes in companies and public authorities, and especially in education, are particularly affected. While most schools and universities worldwide are looking for the correct use of generative AI (Farrelly & Baker, 2023; Holmes & Miao, 2023), companies have

already published reports claiming that this new form of automation will lead to job losses (WEF, 2023; Shrier et al., 2023). Like all AI-based systems, the promise of increased efficiency and effectiveness is particularly attractive to companies. The inadequacies and risks of using such systems, not to mention the disastrous hunger for energy, are often minimised with arguments of innovation and the competitive situation of the free market.

As we have noted, the evidence of discrimination through AI has already been proven several times. Gender or race is the most prominent form (Luccioni et al., 2023; 2024). Here, we see how social injustice in our natural world reaches into the digital space and reinforces it in multiple ways. Therefore, AI can be seen primarily as a mirror of our society rather than a promise of progress to simulate our human intelligence.

The question now arises whether these systems also harbour less obvious cognitive biases and how these can be made visible, like the brilliance bias in LLMs.

Another conceivable, less obvious cognitive bias could be the one we encounter in the real world: people whose brains, contrary to society's expectations, have neurological differences in information processing. This affects around 15 to 20% of our population worldwide (Doyle, 2020). Such conditions primarily include dyslexia, dyscalculia, ADHD and autism spectrum. Current research clearly shows, for example, that there are more than 20 different definitions for dyslexia, which are also culturally and, therefore, socially induced (e.g. Doyle, 2020). The question now arises as to whether certain attitudes, even prejudices, towards these neurodivergent people can also be found in the systems of generative AI. We call the search for such possible distortions **Neuronormative Bias** to make it clear that the perception of these people could be characterised above all by the views of neurotypical people, comparable to the attitudes we find in the real world.

We are further guided by the assumption that text-to-image tools can mainly make such distortions visible. Unlike text-to-text tools (the typical language models), image generation tools (e.g., Stable Diffusion, Midjourney, or Dall•3 or 4, to name only some) process language into visual information and can thus more clearly depict attitudes or prejudices of the actual image than the typical language models. This could be especially true for those who prefer not to read, such as people with dyslexia. While visualisation carries the bias, it makes this bias accessible to people it is biased against. The text-to-image tools do not have the sophistication of linguistic nuance and should, therefore, show much more clearly if, as suspected, cognitive biases towards neurodivergent people are included.

Thus, our working hypothesis is that text-to-image tools contain a neuronormative bias, which is also present in AI-based systems in addition to the apparent gender and ethnicity biases. For the first investigation, we are looking for manifested biases of neurotypical people towards neurodivergent people, using the example of people with Dyslexia.

Let us first look at some common prejudices. These include the fact that dyslexia is often reduced to a weakness or disorder in reading and writing, as well as the prejudice that it manifests itself in a gender-specific way. In addition, we follow the findings of studies that dyslexic people often work in low-paid jobs or are generally underpaid despite the fact of unemployment or homelessness (Maughan et al., 2020; MacDonald et al., 2016; De Beer et al., 2014; Schulte-Körne, 2007). For selecting certain professions, we orientated ourselves on the study's results (Luccioni et al., 2023, 2024), additionally published in a Bloomberg article (Nicoletti & Bass, 2023) in the middle of last year, which selected different high-paid and low-paid job profiles for image generation to prove gender and ethnicity bias. As a basis for generating images of dyslexic people with different high-paid and low-paid job profiles, we selected the job profiles lawyer, caretaker (or social worker) and housekeeper. The other high- and low-paid occupations, as investigated in the Bloomberg article, were disregarded in this first experiment.

A prompting thus revealed that we searched for images of female and male people with dyslexia and additionally carried out a prompting with the corresponding selected occupational categories. Using the tool Gencraft AI, a tool for image generation that promises to create new forms of art, we used it as an app version to generate a first imagination of possible images using a prompt pattern like:

> image of a dyslexic [gender] [profession].

For the first investigation, conducted in December 2023, we used this prompt to generate images of male and female dyslexics who are lawyers, caretakers, or housekeepers, respectively, as shown in Figure 1.

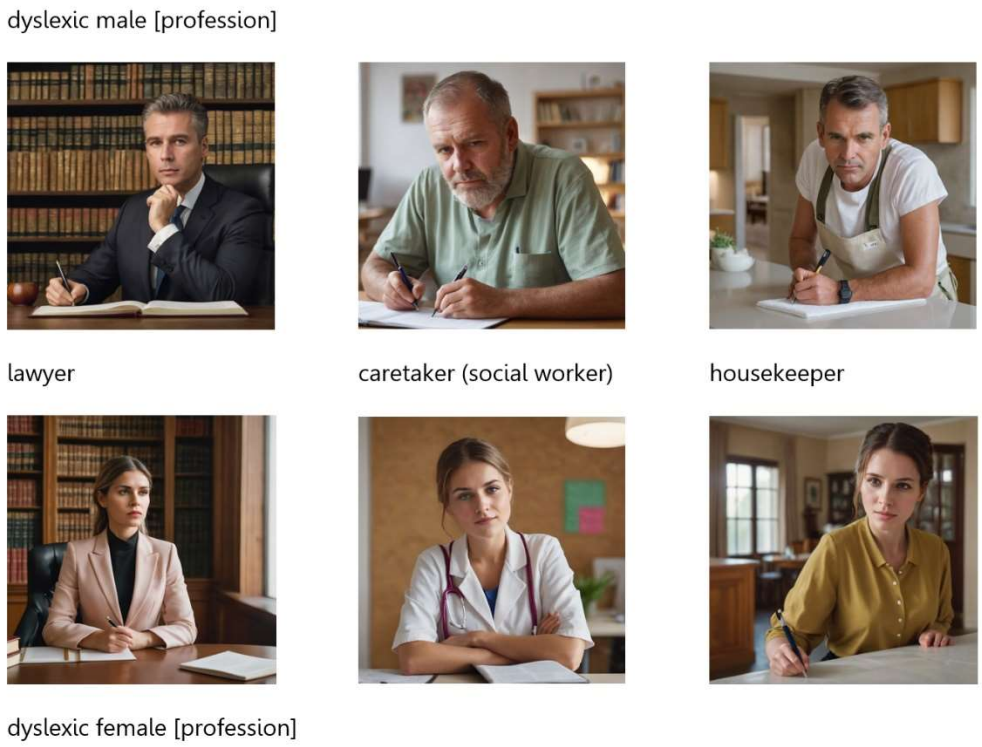

Figure 1: Images of male and female dyslexics with selected professions generated by Gencraft AI (free version, Apple App Store, December 2023)

This first result led us to realise an additional prompt at Stable Diffusion 1.4 where, as in similar tools, a gender and ethnicity bias was already detected (Luccioni et al., 2023, 2024). We wanted to know how this tool represents dyslexic people regarding occupation and gender to detect a possible neuronormative bias, too. For example, the prompt for generating images for dyslexic male and female lawyers compared to male and female lawyers showed a clear result, as Figure 2 shows.

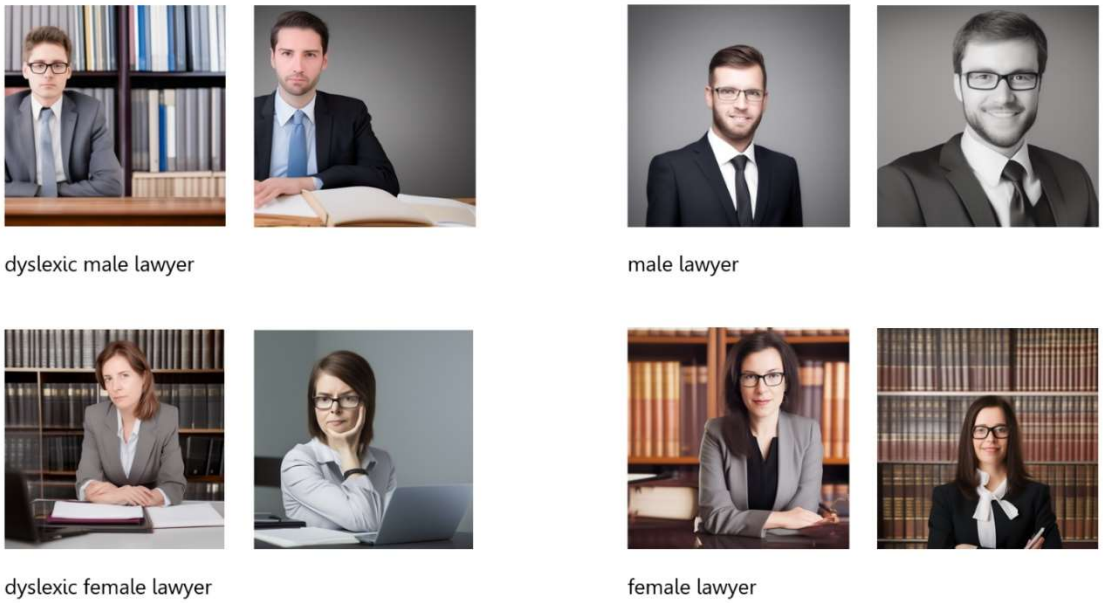

Figure 2: Images of dyslexic and non-dyslexic male or female lawyers generated by Stable Diffusion 1.4 (free version, browser app, December 2023)

These first generated images show a similar neuronormative bias reflecting the prejudices against neurodivergent people that we also encounter in the real world. The dyslexic people appear predominantly earnest, sometimes somewhat distracted or sad, and are often shown with paper and/or pens compared to the group of non-dyslexic human images. This suggests that one of the real-world prejudices was adopted when training the tools. The results are particularly appalling when a comparison is made between neurodivergent and neurotypical images of the different professions. This shows that male and female dyslexics appear happier in low-paid professions and are revealed as unhappier in higher-paid occupations. The images of male and female lawyers, in particular, show this differentiation. Additionally, female dyslexics seem happy in care, social work, and housekeeping compared to their male counterparts' images.

In particular, when we look at the Gencraft AI-generated images, it also becomes apparent that they reinforce our Western view of a beautiful or traditionally attractive person, as they always showed more or less identical faces regardless of profession and neurodivergence. Even though the tools generated images of varying intensity and quality, it became clear that all the images generated were overlaid with a gender (and ethnicity) bias characterised by a strongly reduced understanding of beauty.

For our first experiment, we generated only a few images to investigate our hypothesis of an unconscious bias—the neuronormative bias—towards neurodivergent people. Nevertheless, these first images are shocking, reflecting the aforementioned prejudices of our natural world. Despite the neurodiversity movement that has existed for more than 20 years, these images show alarmingly how little the findings from research have been able to be transferred to reality and how prejudices that have existed for years have only been partially eradicated. These images make it abundantly clear how even less recognised prejudices are reinforced by AI-based systems, further manifesting existing injustices in our society. These systems, therefore, neither contribute to democratising our society nor reduce injustice and social discrimination.

This first observation motivated us to conduct a more thorough investigation, based on Luccioni's research design (Luccioni et al., 2023; 2024), to provide structured evidence of the neuronormative bias using one of the text-to-image tools already analysed for gender and ethnicity. As in the Bloomberg publication (Nicoletti & Bass, 2023) referencing the Luccioni analysis, we will use all seven low- and high-paid occupations. Additionally, we will supplement our investigation with emotion detection to determine a happiness fact, which, according to our hypothesis, will be shown to be gender-specific for people with dyslexia in

the respective characteristics of the occupations. The precise analysis of the neuronormative bias in text-to-image tools is currently being academically analysed elsewhere to document the exact course of the investigation, the defined research design, and the evaluation of the data sets in a scientifically sound manner. This journal article is merely intended as a wake-up call and to show that we encounter non-obvious cognitive biases in using generative AI tools daily, about which we have not yet developed a broad social awareness (in contrast to gender and ethnicity bias). However, as a digital mirror of our values and norms, they show us the path to a genuinely inclusive and neurodiverse society.

Keywords: unconscious bias, neuronormative bias, neurodiversity, generative AI, algorithmisation

References

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp—arXiv *preprint arXiv:2005.14050*.

De Beer, J., Engels, J., Heerkens, Y., & van der Klink, J. (2014). Factors influencing work participation of adults with developmental dyslexia: a systematic review. *BMC public health*, *14*, 1-22.

Doyle, N. (2020). Neurodiversity at work: a biopsychosocial model and the impact on working adults. *British Medical Bulletin*, *135*(1), 108.

Farrelly, T., & Baker, N. (2023). Generative artificial intelligence: Implications and considerations for higher education practice. *Education Sciences*, *13*(11), 1109.

Holmes, W., & Miao, F. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.

Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*.

Luccioni, S., Akiki, C., Mitchell, M., & Jernite, Y. (2024). Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, *36*.

Luccioni, A. S., Jernite, Y., & Strubell, E. (2023). Power hungry processing: Watts driving the cost of a deployment? *arXiv preprint arXiv:2311.16863*.

Lucy, L., & Bamman, D. (2021, June). Gender and representation bias in GPT-3 generated stories in *Proceedings of the third workshop on narrative understanding* (pp. 48–55).

Macdonald, S. J., Deacon, L., & Merchant, J. (2016). 'Too Far Gone': Dyslexia, Homelessness and Pathways into Drug Use and Drug Dependency. *Insights on Learning Disabilities*, *13*(2), 117–134.

Maughan, B., Rutter, M., & Yule, W. (2020). The Isle of Wight studies: the scope and scale of reading difficulties. *Oxford Review of Education*, *46*(4), 429–438.

Nicoletti, L., Bass, D. (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, https://www.bloomberg.com/graphics/2023-generative-ai-bias/, 10 March 2024.

Nishant, R., Schneckenberg, D., & Ravishankar, M. N. (2024). The formal rationality of artificial intelligence-based algorithms and the problem of bias. *Journal of Information Technology*, *39*(1), 19-40, https://doi.org/10.1177/02683962231176842.

Perrigo, B. (2023). Exclusive: OpenAU Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic, *Time Magazine*, 18 January 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/. 10 March 2024.

Schulte-Körne, G. (2007). *Legasthenie und Dyskalkulie: Aktuelle Entwicklungen in Wissenschaft, Schule und Gesellschaft*. Bochum, Germany: Verlag Dr. Winkler.

Shihadeh, J., Ackerman, M., Troske, A., Lawson, N., & Gonzalez, E. (2022, September). Brilliance bias in GPT-3. In *2022 IEEE Global Humanitarian Technology Conference (GHTC)* (pp. 62–69). IEEE.

Shine, I., Whiting, K. (2023). These are the jobs most likely to be lost – and created – because of AI, *World Economic Forum*, https://www.weforum.org/agenda/2023/05/jobs-lost-created-ai-gpt/. 10 March 2024.

Shrier, D.L., Emanuel, J., Harries, M. (2023). Ist Your Job AI Resilient?, Harvard Business Review, Oct 2023, https://hbr.org/2023/10/is-your-job-ai-resilient, 10 March 2023.

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models (arXiv: 2102.02503). arXiv.

Taylor, P. (2023). Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025, https://www.statista.com/statistics/871513/worldwide-data-created/. 10 March 2024.

Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), pp. 433–460, https://doi.org/10.1093/mind/LIX.236.433.