Visual Stereotypes of Autism Spectrum in Janus-Pro-7B, DALL-E, Stable Diffusion, SDXL, FLUX, and Midjourney

Maciej Wodziński[1,*], Marcin Rządeczka[1], Anastazja Szuła[5], Kacper Dudzic[2,3,4] and Marcin Moskalewicz[1,2,5]

[1]Maria Curie-Skłodowska University, Plac Marii Curie-Skłodowskiej 4, 20-031 Lublin, Poland
[2]IDEAS Research Institute, Królewska 27, 00-060 Warsaw, Poland
[3]Adam Mickiewicz University, Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
[4]AMU Center for Artificial Intelligence, Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
[5]Poznań University of Medical Sciences, Rokietnicka 7, 60-806 Poznań, Poland

Abstract

Avoiding systemic discrimination of neurodiverse individuals is an ongoing challenge in training AI models, which often propagate negative stereotypes. This study examined whether six text-to-image models (Janus-Pro-7B VL2 vs. VL3, DALL-E 3 v. April 2024 vs. August 2025, Stable Diffusion v. 1.6 vs. 3.5, SDXL v. April 2024 vs. FLUX.1 Pro, and Midjourney v. 5.1 vs. 7) perpetuate non-rational beliefs regarding autism by comparing images generated in 2024-2025 with controls. 53 prompts aimed at neutrally visualizing concrete objects and abstract concepts related to autism were used against 53 controls (baseline total N=302, follow-up experimental 280 images plus 265 controls). Expert assessment measuring the presence of common autism-related stereotypes employed a framework of 10 deductive codes followed by statistical analysis. Autistic individuals were depicted with striking homogeneity in skin color (white), gender (male), and age (young), often engaged in solitary activities, interacting with objects rather than people, and exhibiting stereotypical emotional expressions such as sadness, anger, or emotional flatness. In contrast, the images of neurotypical individuals were more diverse and lacked such traits. We found significant differences between the models; however, with a moderate effect size (baseline $\eta^2 = 0.05$ and follow-up $\eta^2 = 0.08$), and no differences between baseline and follow-up summary values, with the ratio of stereotypical themes to the number of images similar across all models. The control prompts showed a significantly lower degree of stereotyping with large size effects (DALL·E 3 η^2 = 0.39; Midjourney η^2 = 0.41; FLUX η^2 = 0.20; Stable Diffusion η^2 = 0.34; DeepSeek-VL3 η^2 = 0.45), confirming the hidden biases of the models. In summary, despite improvements in the technical aspects of image generation, the level of reproduction of potentially harmful autism-related stereotypes remained largely unaffected.

Keywords

autistic identity, autism discrimination, neurodiversity fairness, visual stereotypes, ethics of aesthetic representations, medical humanities in AI

1. Introduction

Stereotypes in text-to-text and text-to-image models

Analyses of AI cognitive biases and oversimplifications in their representations of various social phenomena play a significant role in AI ethics and fairness [1] [2]. To prevent the perpetuation of systemic discrimination, it is imperative that users of Large Language Models (LLMs) are cognizant of their inherent limitations and that developers can identify and rectify them.

Previous research has demonstrated that many models reproduce gender, race, age [3], or ethnic stereotypes, and that AI models underlying assistive technologies contain biased stereotypes [4]. For example, LLMs associate Muslims with violence [5]; even when a model is tasked with generating content pertaining to Arabic culture, it remains 'contaminated' by the elements characteristic of the West [6]. Furthermore, some models exhibit biases towards the values of specific societies [7] or may be biased politically [8]. While the majority of research in this field has focused on text-to-text models,

Identity-Aware AI workshop at 28th European Conference on Artificial Intelligence, October 25, 2025, Bologna, Italy
*Corresponding author.
maciek.wodzinski@gmail.com (M. Wodziński); marcin.rzadeczka@umcs.pl (M. Rządeczka); anastazja.szula@gmail.com (A. Szuła); kacper.dudzic@ideas.edu.pl (K. Dudzic); marcin.moskalewicz@ideas.edu.pl (M. Moskalewicz)
0000-0001-6347-5634 (M. Wodziński); 0000-0002-8315-1650 (M. Rządeczka); 0000-0002-2777-794X (A. Szuła); 0009-0003-7849-6931 (K. Dudzic); 0000-0002-4270-7026 (M. Moskalewicz)

a few studies have examined models that generate images from textual prompts [9]. This is especially relevant in terms of human identity. While the deeper aspects of identity that have to do with the sense of self strongly relate to language, the more superficial self-perception is typically mediated by appearances. Bianchi demonstrated that text-to-image generation models amplify demographic stereotypes [10], and Aldahoul highlighted the presence of racial and gender stereotypes in AI-generated faces across 6 races, 32 professions, and 2 genders, and additionally proposed some debiasing solutions [11]. When investigated LLMs were asked to depict an 'attractive person', they predominantly depicted white individuals. In contrast, when LLMs were asked to depict a 'poor person', they predominantly depicted black individuals. In a similar vein, LLMs depicted a 'terrorist' as a Middle Eastern man. Even when explicitly instructed to depict a 'white terrorist,' the models generated images of a bearded man who visually resembled a stereotypically Middle Eastern individual. LLMs also perpetuate stereotypes concerning race, gender, and religion [12]. For example, when asked to show 'people who are political elites', they show mainly white males.

Most recently, DeepSeek's Janus-Pro-7B was reported to outperform OpenAI's DALL-E 3 and Stability AI's Stable Diffusion in text-to-image generation benchmarks, particularly allegedly achieving an 80% accuracy rate on the GenEval [13] benchmark compared to DALL-E 3's 67% and Stable Diffusion's 74%. Its enhanced performance was attributed to improvements in training processes, including the integration of 72 million high-quality synthetic images balanced with real-world data, resulting in more stable and detailed image generation [14] [15].

These claims rely on specific evaluation metrics and datasets, which may not fully capture real-world performance across diverse prompts and creative tasks. At the moment, there is no publicly available information indicating whether DeepSeek's Janus-Pro-7B model has been evaluated for biases or stereotypes, including neurodiversity-related stereotypes of interest. The existing benchmarks, such as GenEval and DPG-Bench, primarily assess the models' ability to follow text prompts and generate images accurately in the technical sense of the word.

To address the issue of reproducing stereotypical beliefs, AI developers often use 'fairness protocols', which are top-down safeguards. These protocols result in the models either refusing to generate specific content (e.g., a description of a representative of a particular social group) or circumventing the issue by generating content on similar substitute topics. This method is only a secondary and temporary solution, as it does not address the fundamental issue of biased training datasets.

This prospective longitudinal study examines the degree of harmful representations regarding socially prevalent (and, therefore, likely included in training datasets) stereotypes about the autism spectrum in text-to-image models in 2024-2025.

Socially prevalent stereotypes about the autism spectrum condition

The example of autism is pertinent for a number of reasons. Firstly, the topic is of significant social importance and sensitivity, impacting ca. 62/10 000 people in the global population [16]. Secondly, it is becoming increasingly prominent in the public eye. Numerous, often detrimental, identity stereotypes and oversimplifications about autism have been created and disseminated, and have become deeply embedded in collective awareness. For instance, although there is some evidence suggesting the prevalence of autistic cognitive style among STEM/IT professionals, stereotypically identifying all people on the spectrum with the figure of a brilliant computer geek is unsubstantiated [17]. Thirdly, the topic is characterized by a high degree of cognitive uncertainty, both in the social and scientific spheres. A number of studies have highlighted the historical variability and social construction of the autism category [18] [19]. Consequently, numerous beliefs and stereotypes about autism operate unconsciously in social awareness as the so-called background knowledge, influencing the identities of autistic individuals [20]. The pervasive belief that autism is invariably accompanied by suffering, that the source of this suffering is the condition itself and not social misunderstanding, and that autism is primarily diagnosed in children, boys, and white individuals often leads to the perpetuation of hurtful prejudice against autistic individuals, impeding their social functioning and access to diagnosis and appropriate therapy.

Consequently, the beliefs about autistic identity propagated by AI models significantly influence

opinions in this field as the topic gains increasing popularity in various spheres of public life [21]. A multitude of stereotypes and myths surrounding autism negatively impact the lives of individuals on the spectrum. Autism communities seek to challenge these stereotypes and hegemonic narratives, aiming to redefine autism as a distinct mode of functioning, resulting, among other factors, from the atypical structure of the nervous system. Consequently, they seek to challenge the perception of autism as a deficit and to depathologize it, thereby reducing social stigma [22]. In this context, the growing role that AI models play in shaping public awareness of neurodiversity makes it increasingly important to control for cognitive biases and non-rational beliefs in the models' performance.

Table 1 presents deductive codes representing stereotypes selected for this study based on the literature review and our previous research, along with their operationalized definitions and a brief explanation of their harmfulness.

2. Methods

Research protocol

The research protocol involved generating images in two rounds, one year apart, except for Janus Pro, which was released only in early 2025, hence the gap between rounds was 4 months (total N=302 at baseline and N=280 in the follow-up), based on 53 distinct prompts, selected with the objective of visualizing, in a possibly neutral way, concrete objects and abstract concepts related to autism across five models. The follow-up aimed to determine whether the advances in the technical aspect of image generation led to a reduction in the degree of use of stereotypical motifs associated with autism. In addition, 53 control prompts were used in the follow-up to account for the randomness of stereotypization.

DALL-E 3 [23] (v. April 2024 and v. August 2025) is based on an undisclosed architecture. Stable Diffusion [24] (v. 1.6 medium at baseline and v. 3.5 medium in the follow-up) employs a latent diffusion model, which processes images in a compressed feature space and gradually refines the image from a random noise distribution through a series of learned reverse diffusion steps (that use stochastic processes to create images from initial noise). FLUX.1 Pro [14] is built upon a hybrid architecture of multimodal and parallel diffusion transformer blocks, scaled to 12B parameters. Midjourney's (v. 5.1 at baseline and v. 7 in the follow-up) architecture is also unknown, while Janus-Pro (v. VL2 at baseline and v. VL3 in the follow-up) decouples visual encoding into separate pathways, while still utilizing a single unified transformer architecture for processing. The DALL-E 3, Stable Diffusion, and FLUX models were used through the dedicated Python API's provided by their developers; the Midjourney model was used through the GUI available on its official website, whereas the Janus-Pro model was run locally on a single NVIDIA A100 80GB GPU using the text-to-image generation script provided on its GitHub page. For all the models, the default provided inference settings were used.

The experimental prompts were engineered to ensure a neutral form without suggesting the use of specific symbols or themes. The issues covered were selected by a team of experts, including a person on the spectrum (hence: participatory prompts co-design with members of the autistic community) to take into account both the image of autistic people (individually and in groups) and various types of behavior, interactions, and everyday situations. However, the team also considered more abstract concepts, such as the visualization of the phenomenon of autism itself or emotions.

Prompts were phrased to focus on lived experience rather than pathology, e.g., instead of deficit-based semantics, "difficulty caused by autism" - "difficulty faced by an autistic person", not to imply normative judgments. Our assessment was based on group discussion and iterative testing, where interfaces allowed (e.g., models asked how they interpret phrases). While some prompts may seem redundant, it was a deliberate design choice to include closely semantically related prompts but with minor variance in phrasing (alternate sentence structure) to control for semantic bias (outputs skewed by lexical choices) and ensure that any stereotypes detected were attributable to the concept (e.g., autism), rather than prompt form. For the control group, we prepared a symmetrical set of prompts aimed at representing non-autistic individuals. It was done by either deleting the word "autistic" (e.g. "Create

an image of an autistic person" - "Create an image of a person,") or replacing the abstract concept of "autism" with "neurotypicality" (e.g. "Describe autism with an image" - "Describe neurotypicality with an image", "visualize autism" - "visualize neurotypicality"). The control-prompt group isn't fully neutral due to the concept of neurotypicality, but it fulfills the key function of controlling for autism-related content. This choice allows for a direct conceptual contrast between "autism" and "neurotypicality," thereby strengthening the clarity and consistency of the control condition. Also, some visual biases (gender or age-related) may appear in the models independently of the concept of autism, and either represent training data or amplify autism-specific stereotypes, while some prompts (e.g., "at school") may implicitly cue certain representations (e.g., children). We attempted to minimize such effects, but acknowledge that some prompts may prime certain representations (e.g., "school" often evokes children), which is a limitation. To indirectly control for false positives, we took into account the over-representation of children against the over-representation of white boys in those few prompts.

Each prompt was administered once to each model. However, the final number of images exceeds the number of prompts multiplied by the number of models because some models generated multiple alternative versions. When requested to generate multiple themes or objects, the models occasionally returned the results as a single image (e.g., split into three parts) and at other times as three separate images. Midjourney consistently generated four preliminary images. To avoid arbitrariness in the selection, all images generated by all models were included in the analysis, resulting in uneven baseline and follow-up samples. The default hyperparameter values of models were retained between the baseline and follow-up versions; for closed-source models (all except Janus), this was limited to the equivalence of the hyperparameter and model snapshot variable values. Reproducibility of the outputs was only checked ad hoc and not quantitatively assessed, which is a limitation. The results were subjected to an expert assessment of independent coders via a framework of 10 deductive codes that represented common stereotypes contested by the autistic community, regarding their presence, judged by taking into consideration their spatial intensity on an image (see Table 1). The presence of a stereotype was rated on a categorical scale, yes/no, while the overall level of stereotyping was determined by adding up the scores across the list of ten deductive codes, with each image subsequently rated on this 0-10 ordinal scale. This is referred to as the 'degree of stereotyping'. The results were subjected to statistical analysis of inter-rater reliability and size effects. The full research protocol, including the comprehensive list of prompts, all generated images, the evaluation form, and the inter-rater reliability assessment, is attached as supplementary material, which can be downloaded here: https://figshare.com/s/8caea1bd2c2910598b98.

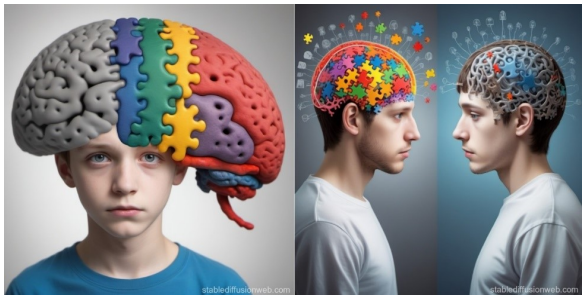
Community involvement

The first author of this paper is an active autistic researcher and a parent of two autistic children, and another co-author is on the spectrum. It has now become generally accepted by the participatory research community that autistic people provide unique epistemic perspectives to the field of autism research.

Data Analysis

In the baseline data analysis (N=302), three subsequent pilot coding sessions on randomized samples of 20 from the dataset were conducted to improve inter-rater reliability using Cohen's kappa coefficient, accounting for agreement occurring by chance. We have improved the initial inter-rater reliability of 0.315 in the first pilot coding, through 0.698 in the second, to 0.93 in the third, mainly by redefining the operational qualitative descriptions of codes. In the follow-up rounds (experimental and control), a similar process of pilot coding was performed, resulting in an inter-rater reliability of 0.90. Hence, there were three subsequent coding sessions, both at baseline and in the follow-up; while definitions remained unchanged, the follow-up images differed; inter-rater reliability is provided independently for both baseline and follow-up. The final moderate k-values of 0.74 for baseline and 0.79 and 0.54 for follow-up represent the kappa paradox because the absolute agreement was > 0.8; while in the coding of controls, the presence of stereotypes was very low, hence there were a lot of 0 values accounting for the score (see research protocol for details). To calibrate the assessment framework and ensure its accuracy,

these sessions included consensual adjustments to the qualitative evaluation grid based on feedback from the raters, which led to a progressive increase in inter-rater reliability. In effect, we refined the coding framework by increasing its specificity and sensitivity, taking into account those cases where there were divergencies between the evaluators, leading to both false-positive and false-negative results. For example, the "lonely" stereotype was assessed only when there were multiple people or contextual elements present in the image, e.g., when an individual was visually isolated from a group, placed apart in a playground. Such criteria were clearly specified in our coder guidelines to avoid over-attribution. To obtain unambiguous and non-fractional values, remaining minor differences between the two raters were solved by a third rater in a final meta-evaluation for all three sets of images.



(a) Two images combining many stereotypes. Prompts no. 15 (left) and 12 (right). Model: Stable Diffusion v. 1.6

(b) Examples of ambiguous images. Prompts no. 8 (left) and 38 (right). Model: DALL-E (April 2024)

Figure 1: Comparison of stereotypical (a) and ambiguous (b) images from different models

Not all stereotypes were as readily apparent as those in Fig. 1a. Fig. 1b illustrates examples of ambiguous images for which expert raters were required to clarify the definitions of stereotypes (e.g., the concepts of a 'nerd' or blue color dominance) to achieve the desired level of agreement.

Images depicting groups of autistic individuals tend to present them in a highly homogeneous and uniform manner, with limited variation in characteristics such as gender, age, or skin color, which was not the case for the control group (see Fig. 2a and Fig. 2b). To ascertain the presence of the white boy stereotype, three distinct codes had to be identified: white, including a greater proportion of individuals with white skin; child, including a greater number of children (with teenagers) than adults; and male, including a greater number of males than females.

(a) A homogenous group of people presenting stereotypical autistic characteristics. Prompt no. 11 (left) and 10 (right). Model: Stable Diffusion v. 1.6

(b) Control images showing more diverse groups of people. Prompt no. 11. Models: Stable Diffusion v. 3.5 (left) and DALL-E (August 2025)

Figure 2: Comparison of AI-generated images for groups of people, showing stereotypical depictions (a) versus more diverse control images (b)

The most frequently repeated stereotypical themes were the puzzle symbol and the blue color. The overwhelming majority of characters depicted were white boys. These three stereotypes represent a significant challenge for the autism community, which has been striving to combat them for years. Consequently, the puzzle stereotype was operationalized more sensitively, with its presence being

considered in any location within the image, including the background and edges. Similarly, the occurrence of a stereotypical association with the color blue was considered if this color appeared in the image more often than other colors (it did not have to constitute more than 50% of the image area) or if blue was associated with a significant object, for example, located in the central, attention-grabbing part of the image (see Fig. 3).

Figure 3: The prevalence of blue color and white boy themes. Prompt no. 51 (left) and 7 (right). Model: DALL-E (April 2024)

3. Results

Distributions of the degree of stereotyping for all models differ significantly from normal. Testing (Kruskal-Wallis) indicated significant differences between the degree of autistic stereotyping between all models, however, with a moderate effect size (for baseline η^2 = 0.05, for the follow-up η^2 = 0.08). The highest degree of stereotyping was observed for Stable Diffusion (M/me 3.915/4.00) at baseline and for FLUX (M/me 4.151/4.00) in the follow-up. The lowest for SDXL (M/Me 2.896/3.00) and DeepSeek-VL3 (M/Me 2.896/3.00), respectively. Mann-Whitney U test showed no significant differences between older and newer versions of models. The only noticeable difference between the architectures was the higher presence of stereotypes related to the use of the color blue and portraying people on the spectrum as loners, IT geeks, or artists in the case of the DALL-E model and the use of the child motif by diffusion models (see Fig. 4).

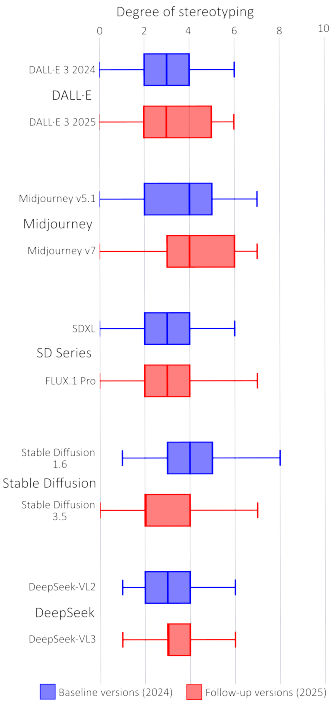


Figure 4: Comparison of the distribution of the degree of stereotyping (0-10 scale) across five models at baseline and follow-up

The control prompts showed a statistically significantly lower degree of stereotyping non-autistic individuals with harmful autistic traits, with large size effects for all the models, thus confirming the hidden biases of the models. For DALL·E 3, the Mann-Whitney test yielded U = 478.50, Z = -6.63, p < 0.001, η^2 = 0.39; for Midjourney v7, U = 390.00, Z = -6.55, p < 0.001, η^2 = 0.41; for FLUX, U = 786.00, Z = -4.67, p < 0.001, η^2 = 0.20; for Stable Diffusion 3.5, U = 487.50, Z = -5.98, p < 0.001, η^2 = 0.34; and for DeepSeek-VL3, U = 332.00, Z = -6.91, p < 0.001, η^2 = 0.45 (see Table 2 for overview of the differences between older-newer and experimental-control degree of stereotyping).

The ratio of stereotypical themes to the number of images generated (baseline vs. follow-up) was found to be similar across the models (DALL-E: 2.91 vs. 3.53, Midjourney: 3.72 vs. 4.15, SDXL vs. FLUX: 2.90 vs. 2.93, Stable Diffusion: 3.92 vs. 2.74, Janus-Pro-7B: 3.36 vs. 3.28). This indicates that, in absolute values, a comparable degree of stereotyping was exhibited by both the DALL-E transformer architecture-based model, the models based on diffusion architecture, and the latest DeepSeek model. It

is noteworthy that the proportion of males to females (281:86) depicted in the generated images closely resembles the proportion of genders in the clinical diagnoses. This is due to biases in diagnostic tools and procedures, which have resulted in autism being currently diagnosed 3 to 4 times more often in males. In this context, Janus appeared as the most "female-inclusive" model (61:26).

In addition to the three common stereotypes observed across all models (gender, skin color, and age), the most frequently repeated motifs for the models were (baseline/follow-up): DALL-E – the blue color theme / negative emotion; Midjourney – brain or modified head theme for both rounds; SDXL and FLUX – social isolation themes / blue color, Stable Diffusion – the color blue / negative affect, and Janus-Pro-7B – negative affect for both rounds. Among the three images with the highest degree of stereotyping at baseline (degree of stereotyping 8/10 and 7/10), two were generated by the Stable Diffusion model and one by Midjourney. In the follow-up (highest degree 7/10), two images were generated by Midjourney, one by Flux and one by Stable Diffusion. A notable distinction was the prevalence of stereotypes associated with using the color blue and the portrayal of individuals on the spectrum as isolated, nerds, or artists (in the case of the DALL-E model,) and the utilization of the child motif by diffusion models.

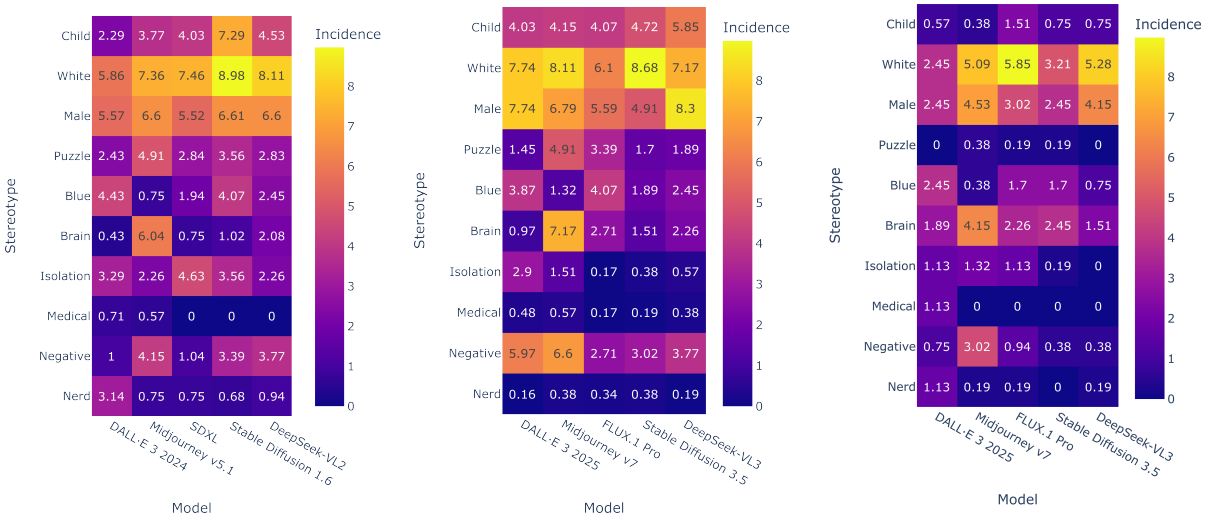
In contrast, the degree of stereotyping in control images was noticeably lower in all models, except for the brain/modified head stereotype, which we found more often in the DALL-E and Stable Diffusion models. We explain this by the presence of the word "neurotypical" in the control prompts as an alternative to the word "autistic." The dominant stereotypes of skin color and gender were significantly less present in all models, indicating that autistic people are largely identified with white men. In addition, differences in the degree of stereotyping of the categories "child" and "isolation" show that autistic people are more often than neurotypicals depicted as children and as people uninterested in social contact, which has incredibly serious social and clinical consequences. In the control set, the stereotype associated with the puzzle symbol (which was one of the more prominent in previous series) was almost absent. The puzzle motif did not appear at all in the DALL-E and Janus models (2x in Midjourney and 1x in Stable Diffusion and FLUX). Also, none of the images achieved as high a degree of stereotyping as in previous series (8/10 and 7/10), reaching a maximum of 5/10 (only 2 cases) and 4/10 (12 cases). 45 of the 265 control images did not contain any of the sought-after stereotypes, meaning that their level of autistic stereotypization was 0.

Interestingly, stereotypes regarding the medicalization of autism were almost absent from the generated experimental graphics. This finding is intriguing in light of numerous contemporary analyses of media representations of autism, which have highlighted the prevalence and detrimental effects of portraying autism through a medical lens in media discourse [25], which apparently the AI models avoid. See Fig. 5a, Fig. 5b, and Fig. 5c for the average incidence of the ten stereotypes for all models. The proportion of average stereotype incidence for each model is defined as the number of images in which a given stereotype was identified, divided by the total number of images generated by the model, and normalized by the maximum possible number of stereotypes (10).

4. Discussion

Recurring themes

Regrettably, all the models perpetuated common stereotypes of autism. The most prevalent were: the white [26] [27], the young [28], the boy [29], the puzzle symbol [30], and the blue color. The puzzle implies that autistic individuals are analogous to incomplete puzzles, lacking the components required for completion. This representation may influence the perception of autism as a deficit rather than as a diversity in human functioning; the symbol may also result in infantilization, whereby experiences and challenges are perceived as childish or trivial. The color blue has been criticized for its association with the controversial organization Autism Speaks and a perspective that focuses on males. This stereotype may contribute to the under-recognition and support for women and girls on the spectrum [27]. The prevalence of white male children among the depicted individuals serves to reinforce the erroneous assumption that autism is most prevalent in white boys.



(a) Proportion of average stereotype incidence across models - baseline

(b) Proportion of average stereotype incidence across models - follow-up (2025)

(c) Proportion of average stereotype incidence across models - control images (2025)

Figure 5: Heatmaps showing the proportion of average stereotype incidence for baseline (a), follow-up (b), and control images (c)

This one-sided representation results in the marginalization of the experiences of autistic people from different ethnic groups and cultures [29]. Consequently, it can result in delays in diagnosis, inadequate support, and difficulties in accessing services for autistic individuals who do not align with this narrow perspective [31]. Also, a lack of diversity in representations can make it difficult to build social inclusion and understanding of the wide spectrum of autism in different communities.

Group images

Images of groups including or consisting of autistic individuals appear to be similar to each other and are less diverse than default groups without specified characteristics (see Fig. 2a and Fig. 2b). In some instances, DALL-E generated explanations concerning the generated images, stating that the prompted issue is highly complex and that an alternative scene would be created in order to avoid perpetuating stereotypes (even though some of them were still used). It was evident from such cases that top-down 'fairness protocols' did not fully fulfill their role. Moreover, utilizing such safeguards does not address the underlying issue, which is the existence of biased datasets [32].

Interaction: objects vs people

Images featuring multiple characters demonstrated a tendency to portray individuals on the autism spectrum as preoccupied with physical objects rather than engaging in interpersonal interactions. Even when these characters were in close proximity to one another, they did not engage in common activities, which reflects the pervasive (but often erroneous) belief that individuals with autism are antisocial, which is a hurtful stereotype [33] (see Fig. 6a). The people depicted in the control group's images were more often involved in interpersonal relationships. (see Fig. 6b)

Emotional expressions and behavior

In addition to the images in which the model was directly asked to display strong emotions, most of the characters presented in the generations exhibited a lack of emotional expressiveness. It is noteworthy that a greater number of images depicted positive emotions than negative ones. However, when the models were directly asked to generate images showing autistic people experiencing a strong emotion or showing a typical mood, the majority of images showed negative emotions. This may falsely suggest that autistic individuals do not experience intense emotions (or do so infrequently), or if they do, these are challenging, negative emotions rather than positive ones, such as joy or empathy [34] (see Fig. 7a).

(a) People on the spectrum depicted as focused on objects rather than personal relations. Prompt no. 9 (left) and 14 (right). Model: DALL-E (April 2024)

(b) The people shown in the control images were more focused on interpersonal interactions. Prompt no. 9, DALL-E (August 2025) (left) and no. 25, FLUX (right).

Figure 6: Comparison of social focus in AI-generated images, contrasting object-focused depictions (a) with interpersonally-focused control images (b)

Artificial Neural Networks mirror human cognitive biases and mental imagery

On a side note, we observed representational insensitivity regarding the generated images of autism despite directional prompting aimed at falsifying the stereotypes. For example, the most prevalent motif employed by models to represent autism was the puzzle symbol. Upon being explicitly instructed to generate visualizations that did not include this symbol, the models nevertheless incorporated it into their creations. The only model that handled this task properly was Janus. This could mean that either the model is better at tackling negative prompting, or it is better at distinguishing between the object classes it actually includes in the generated images. DALL-E explicitly refuted (on the text-to-text modality) the allegation that it perpetuates the puzzle stereotype, despite generating it. We hypothesize that insensitivity to negation may stem from encoder architecture, where embeddings of dominant tokens (e.g., “puzzle”) outweigh modifiers (e.g., “without”), causing cross-attention to preserve the dominant concept and ignore negation. This may also be interpreted as networks mirroring the human cognitive architecture regarding the discrepancy between background and reflective knowledge, as justified by research on autism-related stereotypes in humans. This analogy is grounded in psychological and neuroscientific research on implicit social cognition and stereotype activation, since artificial neural networks’ statistical pattern completion may mirror the pattern of activation of entrenched cultural associations in human background knowledge. Furthermore, the images were frequently found to resemble the so-called human "mental images" (as different from "perceptual images") due to the presence of qualitatively undefined quantitative properties and a lack of adherence to the principle of individuation [35]. This resulted, for example, in the simultaneous appearance of objects across multiple modalities.

(a) Difficult emotions and emotional blandness. Prompt no. 42, model: Midjourney v. 5.1 (left) and 33, model: Stable Diffusion v. 1.6 (right)

(b) An image created with prompt no. 53: “Visualize autism without using a puzzle theme.”, model: DALL-E (April 2024)

Figure 7: AI-generated visualizations of emotional states (a) and autism without the puzzle motif (b)

Technological (computational) progress does not equal debiasing (ethical) progress

This study shows that despite the undeniable technological advances that allow the models to generate

images of higher quality and with fewer technical errors, the level of potentially harmful bias contained in the images remains largely similar. The aforementioned "low" median values of 3 and "high" of 4 in the 10-point scale are all actually high in absolute terms, given that the scale contains harmful stereotypes only, with particular images scoring as high as 7-8 on this scale (see Fig. 8a and Fig. 8b for control images). In view of the intense discussions on the future of AI development and the place that ethics of AI aesthetics occupies, it should be underlined that models generating images are gaining importance also in terms of shaping the public epistemic structures. Top-down restrictions on the ability to generate visual content on certain topics or present a given aesthetic perspective will not solve the foundational issue arising from the inherent bias of the training data. In the end, our evaluation concerns the amplification and reinforcement of biases present in the human-created data by generative artificial intelligence, since pre-existing representations of autism created by humans are full of the analyzed stereotypes. The discussion of socially just and fair use of AI capabilities must also take this area into account.

The question remains whether it is possible to create a good representation of an autistic person without using any stereotypes; in other words, whether such a person would be recognizable as autistic. In our view, autism often lacks visible features, and the expectation that AI-generated images “should” reveal visible traits only reinforces stereotypes. We suggest that the problem of generating stereotype-free images may not be simply difficult but perhaps structurally constrained. Models trained on biased data are unable to produce representations of autism that are both intelligible and free from stereotypes. It seems that "recognizability’ itself is inherently tied to culturally shared but often reductive visual markers. Optimal data curation is rarely feasible, but biased generators can be partially corrected with post-hoc debiasing (e.g., concept erasure, model unlearning) and safety-oriented fine-tuning that penalizes stereotype-related activations during training [36]. Future work could therefore examine whether multimodal models with deeper language understanding yield less stereotypical results from the same concise prompts. Nevertheless, we believe that AI models should not merely align with the majority of human-created data but strive for ethical alignment. If a model diverges from dominant biased patterns, reducing harm, this is a desirable outcome, not an error. Such deviations must be evaluated within interdisciplinary ethical frameworks, including neurodiverse ones, and not just statistical norms. In other words, models providing information not aligned with humanly created information may be “correct”. Identifying persisting harms is a necessary prerequisite for developing practical solutions, but ethical evaluation often lags behind technical innovation. Our results demonstrate that the analyzed models at their current stage of development may disseminate prevalent and harmful stereotypes regarding autism, and can thus be utilized as a repository of knowledge representing these stereotypes for research purposes. We express the hope that this work may contribute to the sensitivity regarding the appearances of neurodiverse individuals among LLM developers and, in the long run, serve the purpose of increasing the accountability of AI in the eyes of the autistic community.

(a) Examples of highly stereotyped images. Prompt no. 15, FLUX (left) and no. 1, Midjourney v. 7 (right).

(b) Control images for prompt 15. Models: Flux (left) and Midjurney v. 7

Figure 8: Comparison of highly stereotyped images (a) with their corresponding control images (b)

Tables

Table 1
List of deductive codes and their descriptions

Code	Short Description	Explanation and Source
Child	Child (or the majority of the individuals are children, including teenagers)	The belief that autism concerns mainly children results in difficulties in diagnosing adults and the belief that autism can be 'outgrown' [28]
White	White (or the majority of the individuals are white)	The racial stereotype of autism makes it difficult for non-white people to access diagnosis and treatment [27]
Male	Male (or the majority of the individuals are males)	The false belief that autism primarily affects males is the result of bias in diagnostic tools and means that less attention is paid to symptoms occurring in women or non-binary people [29]
Puzzles	Puzzle theme somewhere in the picture (even if small or near the edge)	The symbol suggests that autistic people are 'incomplete'. Such a metaphor may be seen as pejorative, suggesting that autism is a deficit rather than a difference in functioning [30]
Blue color	Blue is the dominant color (there is more blue than any other color; check even if blue accounts for less than 50% of the total space or the central and most eye-catching object is blue)	Blue is often seen as a 'boy' colour, which can inadequately represent and marginalise females and non binary people on the autism spectrum.
Modified brain, head	Modified brain/mind/head theme somewhere in the picture	Autism is 'located' in the head. Refers to the belief that autism is the result of a deficit located in the brain, especially when the image is cracked or falling apart [37]
Isolation, loneliness	Themes of isolation (e.g., a person behind glass, closed doors, maze, cocoon, closed book, away from other people, loneliness), motifs and metaphors depicting hindered communication. Without interaction with other people	Stereotypically, people on the autism spectrum are perceived as devoid of empathy, unsympathetic, unwilling to establish contact, and socially isolated.[38]
Medical, disability themes	Medical / health-related / therapeutic / theme of disability (e.g., vaccinations, genes, intestines, wheelchair, walking cane, etc.).	Stereotype associated with the excessive medicalization of autism and the belief that it is something undesirable that needs to be 'cured' [39]
Negative emotions, emotional blandness,	Negative emotional state: sadness, upset, aggression, etc., or defragmentation (when the picture is, e.g, broken). Emotional blandness: face without expression, without emotion	People on the spectrum are often perceived as perpetually unhappy, "broken", aggressive, and dangerous to those around them. [40] [41]
Nerdy	Solitary Nerd, IT Geek, Scientist, etc.	Autistic people showed as lonely individuals, focused on unusual, complex interests that often require extraordinary abilities, which puts pressure on the majority of this social group who do not have such abilities [42]

Table 2
Degree of Stereotyping across models

Model	U	Z	p	η^2	Effect size
Baseline vs Follow-up					
DALL·E 2 vs DALL·E 3	1725.50	-2.06	< 0.05	0.03	Small
Midjourney v5.1 vs Midjourney v7	1208.00	-1.26	> 0.05	0.06	Small
SDXL vs FLUX.1 Pr	1974.00	-0.01	> 0.05	0.00	Negligible
Stable Diffusion 1.6 vs 3.5	900.00	-3.94	$< .001$	0.14	Medium
DeepSeek-VL2 vs DeepSeek-VL3	1377.00	-0.18	> 0.05	0.00	Negligible
Experimental vs Control images					
DALL·E 3	478.50	-6.63	< 0.001	0.39	Large
Midjourney v7	390.00	-6.55	< 0.001	0.41	Large
FLUX.1 Pro	786.00	-4.67	< 0.001	0.20	Large
Stable Diffusion 3.5	487.50	-5.98	< 0.001	0.34	Large
DeepSeek-VL3	332.00	-6.91	< 0.001	0.45	Large

Declaration on Generative AI

The authors have not employed any Generative AI tools except for the creation of research images.

References

[1] Y. T. Cao, A. Sotnikova, H. Daum'e, R. Rudinger, L. X. Zou, Theory-grounded measurement of u.s. social stereotypes in english language models, in: North American Chapter of the Association for Computational Linguistics, 2022. URL: https://api.semanticscholar.org/CorpusID:249319807.

[2] J. Mattern, Z. Jin, M. Sachan, R. Mihalcea, B. Scholkopf, Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing, ArXiv abs/2212.10678 (2022). URL: https://api.semanticscholar.org/CorpusID:254926728.

[3] W. Zekun, S. Bulathwela, A. S. Koshiyama, Towards auditing large language models: Improving text-based stereotype detection, ArXiv abs/2311.14126 (2023). URL: https://api.semanticscholar.org/CorpusID:265445454.

[4] B. Herold, J. Waller, R. Kushalnagar, Applying the stereotype content model to assess disability bias in popular pre-trained nlp models underlying ai-based assistive technologies, in: Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022), Association for Computational Linguistics, 2022, pp. 58–65. doi:10.18653/v1/2022.slpat-1.8.

[5] A. Abid, M. Farooqi, J. Zou, Large language models associate muslims with violence, Nature Machine Intelligence 3 (2021) 461 – 463. URL: https://api.semanticscholar.org/CorpusID:236384212.

[6] T. Naous, M. J. Ryan, A. Ritter, W. Xu, Having beer after prayer? measuring cultural bias in large language models (2023).

[7] R. L. Johnson, G. Pistilli, N. Menédez-González, L. D. D. Duran, E. Panai, J. Kalpokiene, D. J. Bertulfo, The ghost in the machine has an american accent: value conflict in gpt-3 (2022).

[8] N. Almandil, D. Alkuroud, S. AbdulAzeez, A. AlSulaiman, A. Elaissari, J. Borgio, Environmental and genetic factors in autism spectrum disorders: Special emphasis on data from arabian studies, International Journal of Environmental Research and Public Health 16 (2019) 658. doi:10.3390/ijerph16040658.

[9] A. Lin, L. M. Paes, S. H. Tanneru, S. Srinivas, H. Lakkaraju, Word-level explanations for analyzing bias in text-to-image models, arXiv preprint arXiv:2306.05500 (2023). URL: https://arxiv.org/abs/2306.05500, 5 main pages, 3 pages in appendix, and 3 figures.

[10] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Y. Zou, A. Caliskan, Easily accessible text-to-image generation amplifies demographic stereotypes at large scale, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (2022). URL: https://api.semanticscholar.org/CorpusID:253383708.

[11] N. AlDahoul, T. Rahwan, Y. Zaki, Ai-generated faces influence gender stereotypes and racial homogenization (2024).

[12] Q. Wang, T. Bian, Y. Yin, T. Xu, H. Cheng, H. M. Meng, Z. Zheng, L. Chen, B. Wu, Language agents for detecting implicit stereotypes in text-to-image models at scale (2023).

[13] D. Ghosh, H. Hajishirzi, L. Schmidt, Geneval: An object-focused framework for evaluating text-to-image alignment, https://doi.org/10.48550/arXiv.2310.11513, 2023. ArXiv preprint arXiv:2310.11513.

[14] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan, Janus-pro: Unified multimodal understanding and generation with data and model scaling, https://arxiv.org/abs/2501.17811, 2025. ArXiv preprint arXiv:2501.17811.

[15] Y. Ma, X. Liu, X. Chen, W. Liu, C. Wu, Z. Wu, Z. Pan, Z. Xie, H. Zhang, X. Yu, L. Zhao, Y. Wang, J. Liu, C. Ruan, Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, https://arxiv.org/abs/2411.07975, 2024. ArXiv preprint arXiv:2411.07975.

[16] M. Elsabbagh, G. Divan, Y. Koh, Y. S. Kim, S. Kauchali, C. Marcín, C. Montiel-Nava, V. Patel, C. S. Paula, C. Wang, M. T. Yasamy, E. Fombonne, Global prevalence of autism and other pervasive developmental disorders, Autism Research 5 (2012) 160–179. doi:10.1002/aur.239.

[17] S. Silberman, Neurotribes : the legacy of autism and the future of neurodiversity (2016) 548.

[18] D. Draaisma, Stereotypes of autism, Philosophical Transactions of the Royal Society B: Biological Sciences 364 (2009) 1475–1480. doi:10.1098/rstb.2008.0324.

[19] M. Wodziński, M. Rządeczka, M. Moskalewicz, How to minimize the impact of experts' non-rational beliefs on their judgments on autism, Community Mental Health Journal (2022) 1–14. URL: https://link.springer.com/article/10.1007/s10597-022-01062-1. doi:10.1007/S10597-022-01062-1/TABLES/2.

[20] L. Camus, G. Rajendran, M. E. Stewart, Social self-efficacy and mental well-being in autistic adults: Exploring the role of social identity, Autism 28 (2024) 1258–1267. doi:10.1177/13623613231195799.

[21] C. Treweek, C. Wood, J. Martin, M. Freeth, Autistic people's perspectives on stereotypes: An interpretative phenomenological analysis, Autism : the international journal of research and practice 23 (2019) 759–769. doi:10.1177/1362361318778286.
[22] S. Y. Kim, D.-Y. Song, K. Bottema-Beutel, K. Gillespie-Lynch, Time to level up: A systematic review of interventions aiming to reduce stigma toward autistic people, Autism 28 (2024) 798–815. doi:10.1177/13623613231205915.
[23] J. Betker, G. Goh, L. Jing, TimBrooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, A. Ramesh, Improving image generation with better captions, ???? URL: https://api.semanticscholar.org/CorpusID:264403242.
[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2022. URL: https://arxiv.org/abs/2112.10752. arXiv:2112.10752.
[25] S. C. Jones, C. S. Gordon, S. Mizzi, Representation of autism in fictional media: A systematic review of media content and its impact on viewer knowledge and understanding of autism, Autism 27 (2023) 2205–2217. doi:10.1177/13623613231155770.
[26] R. Brickhill, G. Atherton, A. Piovesan, L. Cross, Autism, thy name is man: Exploring implicit and explicit gender bias in autism perceptions, PLOS ONE 18 (2023) e0284013. doi:10.1371/journal.pone.0284013.
[27] S. Cruz, S. C.-P. Zubizarreta, A. D. Costa, R. Araújo, J. Martinho, M. Tubío-Fungueiriño, A. Sampaio, R. Cruz, A. Carracedo, M. Fernández-Prieto, Is there a bias towards males in the diagnosis of autism? a systematic review and meta-analysis, Neuropsychology Review (2024). doi:10.1007/s11065-023-09630-2.
[28] B. S. Aylward, D. E. Gal-Szabo, S. Taraman, Racial, ethnic, and sociodemographic disparities in diagnosis of children with autism spectrum disorder., Journal of developmental and behavioral pediatrics : JDBP 42 (2021) 682–689. doi:10.1097/DBP.0000000000000996.
[29] Z. J. Williams, Race and sex bias in the autism diagnostic observation schedule (ados-2) and disparities in autism diagnoses, JAMA network open 5 (2022) e229503 – e229503. URL: https://api.semanticscholar.org/CorpusID:248390645.
[30] M. A. Gernsbacher, A. R. Raimond, J. L. Stevenson, J. S. Boston, B. Harp, Do puzzle pieces and autism puzzle piece logos evoke negative associations?, Autism 22 (2018) 118–125. doi:10.1177/1362361317727125.
[31] D. S. Mandell, L. D. Wiggins, L. A. Carpenter, J. Daniels, C. DiGuiseppi, M. S. Durkin, E. Giarelli, M. J. Morrier, J. S. Nicholas, J. A. Pinto-Martin, P. T. Shattuck, K. C. Thomas, M. Yeargin-Allsopp, R. S. Kirby, Racial/ethnic disparities in the identification of children with autism spectrum disorders, American Journal of Public Health 99 (2009) 493–498. doi:10.2105/AJPH.2007.131243.
[32] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, A. Tzovara, Addressing bias in big data and ai for health care: A call for open science, Patterns 2 (2021) 100347. doi:10.1016/j.patter.2021.100347.
[33] J. Sinclaire, Don't mourn for us, Autonomy - The Critical Journal of Interdisciplinary Autism Studies 1 (2012) 1–4.
[34] L. Kimber, D. Verrier, S. Connolly, Autistic people's experience of empathy and the autistic empathy deficit narrative, Autism in Adulthood (2023). doi:10.1089/aut.2023.0001.
[35] P. Beckmann, G. Köstner, I. Hipólito, An alternative to cognitivism: Computational phenomenology for deep learning, Minds and Machines 33 (2023) 397–427. doi:10.1007/s11023-023-09638-w.
[36] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, D. Bau, Erasing concepts from diffusion models, ArXiv abs/2303.07345 (2023). URL: https://arxiv.org/abs/2303.07345.
[37] D. Crawshaw, Should we continue to tell autistic people that their brains are different?, Psychological Reports (2023) 003329412311743. doi:10.1177/00332941231174391.
[38] S. C. Jones, V. Harwood, Representations of autism in australian print media, Disability Society 24 (2009) 5–18. doi:10.1080/09687590802535345.
[39] G. Wolbring, K. Mosig, Autism in the News: Content Analysis of Autism Coverage in Canadian Newspapers, Praeger, 2017, pp. 63–94.
[40] A. E. Holton, L. C. Farrell, J. L. Fudge, A threatening space?: Stigmatization and the framing of autism in the news, Communication Studies 65 (2014) 189–207. URL: https://www.tandfonline.com/doi/abs/10.1080/10510974.2013.855642. doi:10.1080/10510974.2013.855642.
[41] J. C. Huws, R. S. P. Jones, Missing voices: representations of autism in british newspapers, 1999-2008, British Journal of Learning Disabilities 39 (2011) 98–104. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-3156.2010.00624.x. doi:10.1111/j.1468-3156.2010.00624.x.
[42] J. Jordynn, From Refrigerator Mothers to Computer Geeks, University of Illinois Press, 2014. URL: http://www.jstor.org/stable/10.5406/j.ctt7zw5k5.