26TH INTERNATIONAL CONFERENCE ON ENGINEERING AND PRODUCT DESIGN EDUCATION
5-6 SEPTEMBER 2024, ASTON UNIVERSITY, BIRMINGHAM, UNITED KINGDOM

UNDERSTANDING HOW UNDERGRADUATE STUDENTS PERCEIVE BIASES IN AI-GENERATED IMAGES, A RESEARCH-THROUGH-DESIGN PROBE

Verónica SILVA[1], Daniel BUZZO[2], Rodrigo HERNÁNDEZ-RAMÍREZ[3] and Hande AYANOGLU[1]
[1]UNIDCOM/IADE, Portugal
[2]CODE University of Applied Sciences, Germany
[3]University of Sydney, Australia

ABSTRACT
The speed and proficiency of generative Artificial Intelligence (AI) systems have proliferated in recent years, enabling more people, including design students, to use AI-generated images for their projects. However, it has been well documented that the Large Language Models supporting AI generators have incorporated troublesome gender, cultural and race biases during training. Undergraduate students, whose visual culture and critical skills are still in development, often lack the capacity to identify such biases in the images they obtain when using AI generators. This can lead to visual outputs that perpetuate prejudiced representations of people. To better understand the nature of this problem and potential ways to mitigate it, we conducted a design probe study on a group of first-semester undergraduate design students in Lisbon, Portugal. The results of this study can be used by teachers to guide their students better and researchers to develop methodologies to help younger generations identify biases in AI generative systems. The impact of this research extends beyond the classroom and can benefit other educators and designers of future AI generative systems. Most importantly, it can contribute to curtailing the perpetuation of race, cultural and gender biases in today's society.

Keywords: AI-Generated images, bias in AI, undergraduate design students, design probes, research through design

1 INTRODUCTION
In the rapidly evolving landscape of AI techniques, the acceleration of ease and quality of automated media generation raises important concerns about the cultural biases we might perpetuate with these systems. The incorporation of AI-powered systems into a growing number of aspects of our lives requires a thoughtful examination of the ethical consequences that the automation of cultural tropes can have. The presence of potentially harmful biases concerning gender, race, age, and cultural background in AI-generated images has been well documented (Naik et al., 2023). These biases perpetuate harmful stereotypes, reinforce societal inequalities, and negatively impact marginalised communities. For instance, AI technology in the justice system demonstrates a clear bias towards black individuals when it comes to detecting and predicting criminals. (Angwin et al., 2016; Malik, 2023). The implications of bias in AI systems can reach beyond the digital interfaces, affecting real-world decisions and perpetuating social injustices.
The field of AI, specifically Generative Machine Learning methods, has undergone considerable progress in recent years, leading to systems with the capacity to produce a broad range of imagery, including hyper realistic images, that can be indistinguishable from photographs. It could be argued that this represents a progression from post-photography, as it challenges conventional photography by exploring the modern image environment filled with abundant images and information (Moreiras, 2023). However, several studies have raised concerns about the unintentional perpetuation of biases often found in AI-generated media. According to Feng et al. (2022), gender biases in image search results significantly influence people's perceptions and reinforce existing biases. In our study, we chose to investigate the views of college students at the undergraduate level, as they are consumers and possible participants in this changing technological environment.

While the creation, reproduction, and perpetuation of cultural bias in imagery have long been subjects of investigation in Sociology, Psychology, and Communication and Media Studies, they remain less explored in the context of AI-aided design education. Accordingly, our research starts with a comprehensive strategy, integrating qualitative and quantitative techniques to investigate how undergraduate students interpret negative stereotypes in AI-generated images. Integrating design probes into the research process enables us to connect theoretical discussions on AI biases with a practical understanding of how young users experience these biases (Gaver, 1999). Many individuals argue that young students may lack the maturity to comprehend and recognise biases, prompting a need to assess the validity of this claim. The results of this research contribute to scholarly discussions and real-world initiatives in the continuous effort toward ethical AI advancement.

2 BACKGROUNDS: BIAS IN GENERATIVE AI

Generative AI, especially deep learning models, has made impressive progress in producing content, particularly images. These mechanisms, powered by extensive datasets, are trained to imitate patterns and styles found within the data used for training. This learning procedure gives rise to concerns regarding the possible perpetuation of biases that exist within the data. Researchers have uncovered a complex connection between culture and gender bias in LLMs, revealing how training in diverse cultural settings may result in varying forms and degrees of biases (Zhou et al., 2023a). Generative AI image models can generate highly detailed and lifelike images but are still susceptible to reproducing and perpetuating societal biases.

AI biases can appear in various ways, such as gender, race, ethnicity, and cultural stereotypes. The most prevalent issue is gender bias, as shown in the research by Gorska et al. (2023), where 76% of images from nine popular text-to-image generators portrayed men. While some may contend that this simply reflects societal gender bias, research has demonstrated the prevalence of gender stereotyping in AI-generated images, as compared to human perceptions, particularly showing a greater degree of gender stereotyping within work-related settings (García-Ull et al., 2023). In general, two main kinds of gender bias have been studied. Representational bias pertains to the unequal portrayal of men and women in different media contexts, especially the over-representation of women in stereotypical feminine positions. Presentational bias involves reinforcing biases in a gender-stereotypical manner, such as women being more inclined to exhibit smiling, calmness, or pitch downward in female-dominated professions (Sun et al., 2023). As AI plays a growing role in shaping decision-making processes and moulding our digital interactions, it is crucial to understand and address these biases.

Nonetheless, there are other forms of biases that demand our attention. Recent studies have shown the existence of age bias in AI systems, indicating the need to also consider the viewpoints of older individuals (Chu et al., 2023). Cave et al. (2020) have also identified the predominance of whiteness in the racialisation of AI and its implications within critical race theory. The creators of these tools already recognise the issue, but resorting to a hasty fix, such as generating random images of people from various ethnic backgrounds, is also not an appropriate solution. Google's Gemini, for instance, was designed to display a variety of individuals but overlooked instances where such diversity should not have been shown, such as women holding the position of pope or black soldiers serving in the German army during WWII (Raghavan, 2024). While altering the databases and historical data used to train AI models may pose challenges, researchers, developers, and policymakers must take proactive measures to tackle these problems.

3 METHODOLOGIES

The study centred on first-semester undergraduate students enrolled in a design program. As newcomers in the industry, these students bring a distinct viewpoint that is not influenced by extensive familiarity with design conventions. This may lead to more genuine responses to AI-generated images. Salminen et al. (2020) state that artificially generated facial images have numerous potential applications, such as creating data-driven personas, advertising, virtual avatars, and fashion. Therefore, students need to understand the biases that may be present in these types of AI-generated images. To maintain the students' interest in the research and gather optimal data, we opted for a design probe study. This approach provides an innovative method to involve young people in immersive environments by using inquiries (Matos et al., 2022). This approach also embraces unpredictability and vagueness, offering a technique for studying specific audience groups during the initial phases of research (Černevičiūtė et al., 2022). Given the potentially sensitive nature of the study's content on biases, it was crucial to

prioritise ethical considerations. Steps were taken to ensure participant anonymity and confidentiality at every stage of the research.

Participants in this research were shown a collection of AI-generated images specifically chosen to demonstrate different types of bias, including gender, race, and cultural stereotypes. In addition, every student was given a Design Probe task package containing a compact journal, cards featuring thought-provoking terms (such as impoverished individuals, substance abusers, migrants, etc.), a paper for composing a letter to the AI tool of their choice, and a set of guidelines. The diary was the central element of the collection, requiring participants to contemplate the four images that Midjourney creates when given a prompt each day for seven days and articulate their emotional response. Prompts were given to assist them in reflecting on and analysing the images, with participants being allowed to express themselves through writing or drawing while ensuring complete anonymity of their responses. This multimodal method sought to capture both overt and covert responses, revealing participants' conscious reactions and potential unconscious perceptions of bias in the AI-generated images.

A significant volume of data was collected through reflections recorded in personal diaries in reaction to biased AI-generated images, written responses on cards to provocative terms, and self-reflections following an attempt at creating a self-portrait with AI. Furthermore, the instructor conducted group discussions with the students to gather their perspectives as well as additional insights that might have yet to be included in their reports. After collecting data, thematic analysis and qualitative coding were used to identify recurring patterns, themes, and variations in the responses from participants. Due to the nature of the topic under investigation, we combined digital and analog materials. Digital probing methods allow for a more exploratory approach to uncovering contextual information and can provide subjective in situ perspectives that are frequently absent in broader studies (Koch et al., 2018; Megarry et al., 2023). This method enabled a comprehensive investigation into how first-semester design undergraduate students perceived, interpreted, and contextualised biases in AI-generated images.

4 INSIGHTS INTO BIAS PERCEPTION AMONG UNDERGRADUATE STUDENTS

The results of the investigation employing design probes with first-semester undergraduate students in design provide fascinating observations into the intricate realm of bias perception in images generated by AI. Contrary to initial assumptions, an important finding revolves around the students' capacity to recognise biases in the images they were shown. These students not only recognised the biases present in the images but also demonstrated an awareness of their own biases or those prevalent in society, aligning closely with Perry et al.'s (2015) proposal. The students who had no awareness of biases in the images also exhibited immature behaviour throughout the semester. It is possible that their limited experience directly impairs their ability to recognise both their own biases and those of others.

We gathered information from 25 students, seven male and 18 female, between the ages of 18 and 20. In analysing the data collected from the students' reactions to bias in AI-generated images (Table 1), several notable patterns emerged. The dataset contains the reactions of the students within a span of seven days, during which they evaluated one set of images generated with Midjourney per day. The first column indicates their gender, feminine or masculine, and each daily evaluation includes a binary indicator of whether the student perceived bias (left column) and a coded summary of their comments (right column). The final column aggregates the number of instances out of seven where each student identified bias. Notably, the data reveals that a third of the students—highlighted in red—either did not perceive any bias or identified it in only one instance, whilst more than a third perceived bias in 5 or more of the images. This suggests a varying degree of sensitivity to bias among the students, with some potentially lacking the critical awareness or tools needed to recognize biases in AI-generated images. While this consideration is significant for design students, it is imperative to cultivate a comprehensive understanding of how students, in general, respond to bias. This knowledge is essential for educating future generations with heightened awareness, ultimately contributing to the reduction of bias in society.

Upon coding the responses to the AI-generated images, we identified nine distinct types of reactions to perceived biases. These codes are as follows: Feminism (F), Racism (R), Stereotype (S), Underrepresentation (U), Pity (Y), Image Quality (IQ), Props (P), Looks (L) and Other (O). Notably, students who perceived bias frequently commented on gender (F) and racial (R) biases, highlighting issues of stereotypes (S) and underrepresentation (U). In contrast, those who did not perceive biases tended to focus on the aesthetics of the images, discussing aspects such as image quality (IQ), the props used in the background (P), and the physical appearance or looks (L) of the AI-generated humans. This

differentiation in focus underlines the varied perspectives among students and suggests that those who are more attuned to bias are inclined to critique the underlying social implications, while others concentrate on superficial or technical elements of the images.

Table 1. Data that shows students' reactions to bias in AI-generated images

Gender	Day 1		Day 2		Day 3		Day 4		Day 5		Day 6		Day 7		Total
F	Yes	S	No	IQ	Yes	F	Yes	R, S	Yes	F, R	Yes	F	No	O	5/7
M	No	O	No	L	No	O	No	Y	No	L	No	L	Yes	O	1/7
F	Yes	F, R	Yes	U	Yes	F, R	Yes	R	Yes	F, R	Yes	F	No	P	6/7
F	Yes	F	Yes	U	Yes	F	Yes	R, S	Yes	L	No	L	No	S	5/7
F	Yes	F	No	Y	Yes	F	No	Y	Yes	F, R	Yes	R	Yes	R	5/7
M	N/A														N/A
F	Yes	F	Yes	S	Yes	F	Yes	R, S	Yes	L	Yes	F	Yes	R	7/7
F	No	IQ	No	IQ	No	IQ	No	IQ	No	IQ	No	IQ	No	IQ	0
F	Yes	F, R	No	Y	Yes	F, R	Yes	R, S	Yes	S	No	L	No	P	4/7
M	No	IQ	No	IQ	No	IQ	No	O	No	IQ	No	IQ	Yes	O	1/7
M	Yes	F, R	Yes	S	Yes	F, S	Yes	R	Yes	F, R	Yes	U	Yes	R	7/7
F	Yes	F, R	Yes	S	Yes	F, R	Yes	R, S	Yes	F	Yes	U	Yes	R, S	7/7
F	Yes	F	Yes	S	Yes	F	Yes	S	Yes	S	Yes	S	Yes	S	7/7
M	No	O	No	IQ	No	IQ	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
F	No	IQ	No	Y	No	P	No	O	No	L	No	IQ	Yes	R, S	1/7
M	Yes	F	No	IQ	Yes	F	Yes	R, S	Yes	F, R	Yes	R	No	IQ	5/7
F	Yes	F	Yes	S	No	P	No	IQ	No	IQ	No	IQ	No	P	2/7
F	Yes	F	Yes	R	Yes	S	Yes	R	No	O	Yes	F	No	P	5/7
F	No	IQ	No	IQ	Yes	F	No	IQ	Yes	F, R	No	IQ	Yes	R, S	3/7
F	No	IQ	No	Y	No	IQ	No	Y	Yes	F	Yes	F	Yes	R, S	3/7
F	No	IQ	No	IQ	No	IQ	No	L	No	IQ	No	IQ	N/A	IQ	0
F	No	IQ	No	Y	No	IQ	No	Y	No	L	N/A	N/A	No	IQ	0
F	Yes	F, R	Yes	R, S	Yes	F, S	Yes	R, S	Yes	F, R	Yes	R	Yes	R	7/7
F	N/A														N/A
M	No	IQ	No	Y	Yes	P	No	L	No	L	No	L	No	P	1/7

In general, the images that students found to be most biased were those depicting gender representation, according to Figures 1 and 2. This perception was predominantly observed by female students, while male students exhibited the lowest level of awareness towards biases present in the study. However, this idea is familiar, as fMRI research has demonstrated that individuals process information from members of their own group and those outside it in distinct ways, impacting how they perceive and interact with ingroup and outgroup members (Molenberghs et al., 2018). It is natural for students to encounter challenges when recognising biases from social groups that are different from their own to some degree. Moreover, some students who expressed negative views towards at-risk communities, such as immigrants, remained oblivious to any prejudices in the images presented to them, thereby confirming the assertion made earlier.

A notable discovery was that students are reluctant to use AI to generate images. We received feedback such as "You can perceive that AI has a lot of stereotypes," "Everyone looks the same," etc. They argue that the images contain numerous imperfections, fail to depict reality accurately, are presented without proper context, and perpetuate stereotypes and preconceptions. Once more, this appears to be a prevalent response observed in individuals, as they often exhibit heightened sensitivity towards AI's inconsistent performance and tend to make harsher judgments regarding the fairness of AI decisions compared to their assessments of human experts (Jones-Jang et al., 20220), as well as a general preference for human-made art or images than AI-generated (Zhou et al., 2023b). However, they will encounter AI-generated images, so it is crucial for them to grasp the prejudices that may be present in such images.

Figures 1. and 2. Images created with Midjourney on December 2023.
Prompts: A doctor (left), CEO (right)

5 CONCLUSIONS

The results from a study using design probes showed that new undergraduate students in a design course were able to identify biases in the AI-generated images they were shown. Contrary to the initial expectations, most students not only recognised biases but also demonstrated an understanding of their own biases and those that are common in society. Female students, especially, were noted to be more sensitive to gender representation biases in comparison with their male classmates. The study emphasised the significance of taking into account some students' limited exposure and potential lack of maturity in identifying biases, underscoring the necessity for enhanced support in educational environments. The research also revealed that students were hesitant to utilise AI for producing images due to worries about flaws, absence of context, and reinforcement of stereotypes. Overall, the study provides significant findings that can assist educators in supporting students and shaping the design of approaches to increase understanding of biases in AI-generated content.

In the future, we aim to replicate this study with students from various classes. This will allow us to gather a greater volume of data and enhance the quality of the study. As anticipated, some students did not fully complete the exercise, resulting in our analysis being based only on complete data. This study is part of a PhD research on AI bias. The data collected in this experiment will serve as the foundation for developing a model to instruct present and future designers about bias in AI-generated media. By fostering a new wave of designers with advanced technical skills and deep ethical awareness, we can lay the groundwork for a future in which AI technologies are created, analysed, and employed with an increased emphasis on accountability and diversity.

ACKNOWLEDGEMENTS

The study was supported by UNIDCOM under a grant from the Fundação para a Ciência e Tecnologia (FCT) No. UIDB/00711/2020 attributed to UNIDCOM – Unidade de Investigação em Design e Comunicação, Lisbon, Portugal. The research project is funded by FCT with reference no. 2023.02281.BD.

REFERENCES

[1] Angwin J., Larson J., Kirchner L. and Mattu S. (2016, May 23). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Cave S. and Dihal K. (2020). The Whiteness of AI. Philosophy & Technology, 33(4), pp.685–703. https://doi.org/10.1007/s13347-020-00415-6.

[3] Černevičiūtė J. and Liebutė L. (2022). CULTURAL PROBES METHOD IN DESIGN RESEARCH: CREATIVITY IN SKETCHES. Creativity Studies, 15(1), pp.169–181. https://doi.org/10.3846/cs.2022.15473.

[4] Chu C. H., Donato-Woodger S., Khan S. S., Nyrup R., Leslie K., Lyn A., Shi T., Bianchi A., Rahimi S. A. and Grenier A. (2023). Age-related bias and artificial intelligence: A scoping

review. Humanities and Social Sciences Communications. https://doi.org/10.1057/s41599-023-01999-y.
[5] Feng Y. and Shah C. (2022). Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), pp.11882–11890. https://doi.org/10.1609/aaai.v36i11.21445.
[6] García-Ull F.-J. and Melero-Lázaro M. (2023). Gender stereotypes in AI-generated images. El Profesional de La Información, e320505. https://doi.org/10.3145/epi.2023.sep.05.
[7] Gaver B., Dunne T. and Pacenti E. (1999). Design: Cultural probes. *Interactions*, *6*(1), pp.21–29. https://doi.org/10.1145/291224.291235.
[8] Gorska A. M. and Jemielniak D. (2023). The invisible women: Uncovering gender bias in AI-generated images of professionals. Feminist Media Studies, 23(8), pp.4370–4375. https://doi.org/10.1080/14680777.2023.2263659
[9] Koch D. and Maaß S. (2018). Digital Probes Kit: A Concept for Digital Probes. I-Com, 17(2), pp.169–178. https://doi.org/10.1515/icom-2018-0016.
[10] Jones-Jang S. M. and Park Y. J. (2022). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. Journal of Computer-Mediated Communication, 28(1), zmac029. https://doi.org/10.1093/jcmc/zmac029.
[11] Malik S. (2023, November 28). *Artificial Intelligence and racial justice in the criminal system*. HRRC. https://www.humanrightsresearch.org/post/artificial-intelligence-and-racial-justice-in-the-criminal-system.
[12] Matos S., Silva A. R., Sousa D., Picanço A., Amorim I. R., Ashby S., Gabriel R. and Arroz A. M. (2022). Cultural probes for environmental education: Designing learning materials to engage children and teenagers with local biodiversity. PLOS ONE, 17(2), e0262853. https://doi.org/10.1371/journal.pone.0262853.
[13] Megarry J., Mitchell P., Rittenbruch M., Kao Y., Christensen B. and Foth M. (2023). Probing for Privacy: A Digital Design Method to Support Reflection of Situated Geoprivacy and Trust. Digital Society, 2(3), 55. https://doi.org/10.1007/s44206-023-00083-x.
[14] Molenberghs P. and Louis W. R. (2018). Insights From fMRI Studies into Ingroup Bias. Frontiers in Psychology, 9, 1868. https://doi.org/10.3389/fpsyg.2018.01868.
[15] Moreiras C. (2017). Joan Fontcuberta: Post-photography and the spectral image of saturation. *Journal of Spanish Cultural Studies*, *18*(1), pp.57–77. https://doi.org/10.1080/14636204.2016.1274496.
[16] Naik R. and Nushi B. (2023). Social Biases through the Text-to-Image Generation Lens. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pp.786–808. https://doi.org/10.1145/3600211.3604711.
[17] Perry S. P., Murphy M. C. and Dovidio J. F. (2015). Modern prejudice: Subtle, but unconscious? The role of Bias Awareness in Whites' perceptions of personal and others' biases. Journal of Experimental Social Psychology, 61, pp.64–78. https://doi.org/10.1016/j.jesp.2015.06.007
[18] Raghavan P. (2024, February 23). *Gemini image generation got it wrong. we'll do better.* Google. https://blog.google/products/gemini/gemini-image-generation-issue/.
[19] Salminen J., Jung S., Chowdhury S. and Jansen B. J. (2020). Analyzing Demographic Bias in Artificially Generated Facial Pictures. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–8. https://doi.org/10.1145/3334480.3382791.
[20] Sun L., Wei M., Sun Y., Suh Y. J., Shen L. and Yang S. (2023). Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image Generative AI. https://doi.org/10.48550/ARXIV.2305.10566.
[21] Zhou K. Z. and Sanfilippo M. R. (2023a). Public Perceptions of Gender Bias in Large Language Models: Cases of ChatGPT and Ernie. https://doi.org/10.48550/ARXIV.2309.09120.
[22] Zhou Y. and Kawabata H. (2023b). Eyes can tell: Assessment of implicit attitudes toward AI art. I-Perception, 14(5), 20416695231209846. https://doi.org/10.1177/20416695231209846.