CulturAI: Informing Cultural Alignment for Text-to-Image Generators

Sarah Kiden*
Tayyaba Azim*
University of Southampton
Southampton, UK
sk3r24@soton.ac.uk
ta7g21@soton.ac.uk

Maira Klyshbekova*
Oriane Peter*
King's College London
London, UK
maira.klyshbekova@kcl.ac.uk
oriane.peter@kcl.ac.uk

Gisela Reyes Cruz*
Damian Eke*
Aislinn Bergin*
Maria Waheed*
University of Nottingham
Nottingham, UK
gisela.reyescruz@nottingham.ac.uk
damian.eke@nottingham.ac.uk
aislinn.bergin@nottingham.ac.uk
maria.waheed@nottingham.ac.uk

Abstract

Text-to-Image (T2I) generators and models are increasingly becoming popular across the world, and are being used to generate images and illustrations for various purposes including education, marketing, media and as sources of inspiration. While they offer promise, evidence suggests that T2I models present challenges related to over-representation of some cultures (e.g. mostly Western), harms to minoritised cultures, and geographical erasure. We report on an ongoing project, *CulturAI*, investigating the state of cultural representation in T2I models, and developing criteria and quantitative and qualitative evaluation metrics to inform cultural alignment for T2I generators. Drawing on a scoping review of literature on the topic, and three virtual workshops with 33 experts and individuals from diverse cultural backgrounds, initial insights indicate the complexity of evaluating cultural sensitivity of T2I models and the need to expand the discussion to include experts on culture and people with lived experiences.

(Wordcount = 2410)

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; *HCI design and evaluation methods*; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Text-to-Image Modelling, AI Audit, Cultural Representation, Generative Artificial Intelligence, Responsible AI

ACM Reference Format:

Sarah Kiden, Tayyaba Azim, Maira Klyshbekova, Oriane Peter, Gisela Reyes Cruz, Damian Eke, Aislinn Bergin, and Maria Waheed. 2025. CulturAI: Informing Cultural Alignment for Text-to-Image Generators. In *Proceedings of UK AI Research Symposium (UKAIRS 25).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

*Equal contribution by authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UKAIRS 25, Newcastle, UK

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Text-to-Image (T2I) models such as DALL-E, Midjourney and Gemini are gaining popularity across global communities, where they are used to generate visual representations of a variety of images and illustrations. Whilst they offer flexibility and lower the access barrier for image making, evidence suggests that, when prompted without a specific cultural identity, the outputs of T2I models frequently reflect Western cultural norms and perspectives, and may perpetuate harms, misinformation and cultural or geographical minimisation [6, 8, 10, 18]. Although recent research has focused on the biases and stereotypes these models can reinforce (e.g., misrepresentation, erasure, stereotyping), there has been less attention on how well these models capture the full diversity and complexity of different cultures. Moreover, 'culture' as a construct is complex to define due to its multifaceted nature [14]. To navigate this, Artificial Intelligence (AI) research often conceptualises culture through discernible patterns like skin tone or facial features, or by utilising cultural proxies. These proxies are essentially measurable stand-ins for the broader concept of culture, enabling researchers to benchmark how AI models grasp cultural aspects [7, 19]. Cultural proxies include, but are not limited to, demographic factors (e.g., gender, name, geography), semantic material attributes (like apparel, cuisine, cultural artefacts), and semantic non-material attributes (such as beliefs, norms, games, myths, events, celebrations).

In this project, we are interested in examining the missing perspectives on how culture is represented and evaluated in AI-generated images. This paper reports on an ongoing project, *CulturAI: Informing Cultural Alignment for Text-to-Image (T2I) Generators*, which is a Responsible AI UK (RAi UK) funded Coordinating Keystone project, seeking to answer these research questions:

- **RQ1**: What is the current state of cultural representation in T2I models?
- **RQ2**: How can we leverage participatory approaches to measure cultural representation of T2I models?
- **RQ3**: How can we integrate cultural considerations into the evaluation of T2I models?

- **RQ4**: How can we ensure that the evaluation of T2I models is representative of more culturally-inclusive perspectives?

2 Background

A fundamental requirement for the usability of Generative AI models is 'AI alignment' – the process of steering these models' outputs to align with what (some) humans prefer. This steering mechanism is crucial for ensuring the models generate relevant and desired results [16]. However, the preferences currently shaping this alignment largely reflect those of Western, Educated, Industrialised, Rich, and Democratic (WEIRD) societies [2, 9]. Despite being considered "standard subjects" in some HCI research, WEIRD communities are not representative of the broader population in the world [9, 20]. Consequently, there is a growing effort to develop more inclusive and representative approaches to alignment [13]. A crucial step in steering this evolution of alignment is to first identify and understand the current misrepresentation in these models. More recently, these investigations have expanded from textual models to encompass T2I models, with a growing effort to quantify the misrepresentation, neglect, or stereotypical depiction of certain cultures. [3, 11, 12, 21]. These studies often rely on benchmark datasets to compare and assess these models. For instance, Zhang et al. [21] crawled Google Chrome to obtain images of nine cultural objects from 30 countries, selecting items that are universal and easily recognisable. Similarly, Kannen et al. [12] extracted approximately 300,000 cultural artefacts from WikiData. They selected artefacts which are "commonly held" across a range of cultures, using geographical location as a proxy for culture. Such approaches have been criticised for lacking community consultation and adopting a "technosolutionist and positivist" view of representation, treating it as objective and quantifiable [17].

In contrast, Qadri et al. [17] argue that representation is not a static concept that can be objectively evaluated, but rather an ongoing interpretive process that is contested and negotiated. They propose a "*thick evaluation*" approach to cultural representation, which involves capturing in-depth, qualitative reflections and discussions among participants as they interpret and evaluate AI-generated images. Instead of relying on predetermined, static features to measure cultural representation, their approach focuses on the contextual, discursive, and mutually negotiated nature of image interpretation. Inspired by Qadri et al.'s thick evaluations [17], we conducted stakeholder engagement sessions where diverse participants generated and evaluated images, then discussed visual representations of T2I models in light of their cultural background.

3 Study Design & Research Status

We conducted three virtual workshops between March and April 2025. Our participant inclusion criteria included individuals older than 18 years and who are either identified as an expert in the topics related to the project (i.e., cultural studies practitioner, professional academic, AI researcher, policymaker, etc.), or are current/potential users of T2I models with culturally diverse backgrounds (e.g., individuals or groups who bring diverse perspectives shaped by different cultural backgrounds and experiences, belonging to underserved communities). This criterion ensures a heterogeneous population capable of navigating the strengths of AI systems from a multicultural lens and sharing insights that are often underrepresented in mainstream discussions. Participants were recruited through direct outreach via email, known networks, and personal channels on LinkedIn. A total of 33 participants joined the sessions, 28 self-identified as experts in the first two workshops, while 5 as general public in the third workshop. Participants joined the workshops from different countries including the UK (where the research team is based), Bangladesh, India, Kenya, Namibia, Nigeria, South Africa, Switzerland, and USA, among others. In designing our workshops, we followed the general participatory workshop and focus group guidelines [5], and used the Miro Software [1] as our virtual interactive board. The participants were first presented an overview of the project and findings of a scoping review [15] of the literature on T2I models, cultural alignment, and T2I benchmarks and evaluations. Each workshop was divided into three parts:

(1) Participants worked in small groups on a Miro virtual board [1] to complete activities that made them reflect on the context of use of AI-generated images, from both their personal use experience and AI images they have encountered elsewhere. They were further asked to reflect on cultural representation in those use cases. Then, each group collectively generated images using the Gemini multimodal app [1] using prompts adapted from the Mad Libs Hall Pass [2] game, in which cultural elements were iteratively added. For instance, '`generate an image of X`', and then '`generate an image of X, with/in Y, Z` and A', and so on. Gemini was preferred over other T2I generators because it did not require a paid subscription to generate numerous images.
(2) Participants were presented with a general definition of culture and a long list of cultural elements discussed or evaluated to different extents in AI research. These elements were identified from the scoping review conducted by the research team prior to the workshops and were grouped in three clusters: 1) demographic proxies, 2) semantic material proxies, and 3) semantic non-material proxies. After spending some time reading this information, participants were asked to suggest any missing cultural elements in the literature while discussing their reflections with the rest of the attendees and facilitators. Some of the elements identified by participants are shown in Figure 1.
(3) Participants were asked to reflect on how cultural sensitivity in AI-generated images can be evaluated, by proposing and discussing criteria and evaluation metrics for conducting a comparative analysis of T2I outputs in the future.

4 Initial Insights

4.1 Widespread and Diverse use of T2I

Most of the participants reported having used a T2I model for a wide range of use cases. These included education and research (illustrations, demonstrations, and images for presentations), social media, content creation, scenario planning, brainstorming or ideation (e.g., traditional art, tattoos, logos, film or online/board game characters), and marketing materials (posters or digital ads).

[1] Gemini interface https://gemini.google.com/app
[2] Word game where a player prompts people for words to substitute blank spaces https://madlibs.com/. In our case, the blank spaces were X, Y, Z and A.



Figure 1: Miro board [1] showing the results of Activity 2. Each pink square is an element or sub-element of culture identified by participants in one of the workshops.

4.2 WEIRD by Default

The major concern participants raised during their previous usage centred around the lack of representation outside of WEIRD societies [9]. Beyond the under-representation of non-WEIRD cultures, participants noted the perpetuation of a colonial gaze and the cultural erasure of groups, such as Black Women in the representation of Woman's Rights struggles. Such concerns were reiterated during the iterative prompting of Gemini. As observed by participants, non-specific prompts (i.e., '`generate an image of X`') tended to present WEIRD-centric-looking images. Iteratively adding a description would reveal the implicit assumption embedded in the model, see figure 2. Such results echo previous research, exploring the perspectives of people from cultures that are misrepresented in AI generation [13, 18, 19].



Figure 2: Images generated using the prompt 'generate an image of a woman's hair' (left), then iterated to include 'African woman' (right)

4.3 Missing Cultures

Beyond the issues of WEIRD as a default, an initial finding of the workshop was that some cultures could not be represented at all. One participant reported that during their previous usages of T2I models, they could not generate images of their Rohingya (a stateless Indo-Aryan ethnolinguistic group) culture, even when using "*very specific instructions and very specific descriptions*". This impression was confirmed during the image generation exercise, in which the participant never felt represented in the produced images. Notably, another participant of the group pointed out that some images we generated related to Rohingya culture actually depicted scenes from Bangladesh (Figure 3). Both participants theorized that this might be attributed to the significant presence of Rohingya refugees in Bangladeshi camps, which could have influenced the model's training data and semantic representation.



Figure 3: Gemini's representation of Rohingya and Bangladeshi children's in school respectively

4.4 Necessity of Culturally Specific Expertise

Across the three workshops, a recurring theme was that some cultural elements are difficult to identify if you are not part of the represented culture. Some of the implicit assumptions of the model could only be identified through nuanced and localised cultural sensitivities. For example, a participant attempting to generate an image of a North Eastern Indian breakfast (Figure 4) found that multiple prompting iterations were necessary, and the results still largely represented an upper-class demographic, which was not the case with some more dominant cultures. The participant's contextual knowledge of Indian cuisine was essential in identifying these biases, which could otherwise have gone unnoticed. Similarly, the presence of members of both Rohingya and Bangladeshi cultures was essential in recognising and understanding the misleading representation of the Rohingya culture. The discussions of participants among themselves during the workshop underscored the importance of combining specialised perspectives. They also pointed to the role of discursive practices in detecting and evaluating cultural representation. The necessity to include a variety of perspectives was also underscored by the participants. For example, one person noted that:

"*it's hard actually to evaluate the cultural sensitivity given that even the way of evaluating cultural sensitivity is different for a different culture and different people. So the best thing and the best concept I can see is the representations of people from different cultures during the evaluation.*"



Figure 4: Images generated iteratively using increasingly specific prompts. Each prompt used is indicated underneath the corresponding image.

4.5 Quantifiable and Unquantifiable Culture

One of the aims of the study was to develop criteria and evaluation metrics of T2I models in a participatory bottom-up approach. Participants identified both quantifiable and unquantifiable criteria and evaluation metrics. These include accuracy, correctness (historic events and representation), realism or quality of images, demographic features (e.g., body shape, skin tones, hair), precision, presence/absence of elements (checking yes/no), lists, numeric ranking/rating (Likert scale - 1 to 10), percentages, authenticity, unique voice and style, level of detail, backgrounds or image surroundings, and symbols, amongst others. However, consistent across the workshops was that some cultural elements (e.g., demographic features and cultural artifacts) cannot be quantitively and numerically measured or ranked. Participants described some images as "*feeling off*", not seeming "*natural or genuine*", or making them "*feel sad*" in cases where their cultures were severely misrepresented. These assessments were based more on emotions, atmosphere or overall 'vibes' of the outputs, and were dependent on each participant's cultural perceptions and experiences. Participants recommended that in such cases, T2I evaluators should review the image outputs in context and with guidance from the cultural experts or people with a living experience in specific cultures. One participant stated that:

"*I think the metrics that you can use, because we are dealing with culture and we're dealing with heritage, I think it's something that is more people-centred. For us to know the performance, I think it's better to go back to the communities because heritage is community-based, so we can only know whether our system is working well by hearing from the communities.*"

5 Future Work

Our initial workshops have already explored participatory approaches for measuring cultural representation (RQ2) and generated insights that will now guide our investigation into the remaining questions. Specifically, these findings will inform the design of methodologies to understand the current state of cultural representation in T2I models (RQ1). Our immediate next step involves a thematic analysis [4] of workshop transcripts, which will be validated through a follow-up consultation survey with participants.

Building on these collective insights, we will design and explore testing methodologies that integrate cultural considerations, are inclusive, and leverage participatory approaches (RQ2, RQ3, RQ4). The intended outcome of this work is an interactive website illustrating the representation of specific cultures in T2I models over time. This platform aims to educate a broad audience (general public, policymakers, researchers, etc.) on the cultural sensitivities of AI-generated images. Furthermore, this research seeks to guide future investigations into the development of improved cultural alignment strategies for T2I generation.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council "AI UK: Creating an International Ecosystem for Responsible AI Research and Innovation" Award: EP/Y009800/1.

References

[1] 2024. *Miro Visual Workspace.* https://miro.com

[2] Mohammad Atari, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. 2023. Which Humans? doi:10.31234/osf.io/5b26t psyarxiv:10.31234

[3] Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5136–5147.

[4] Virginia Braun and Victoria Clarke. 2021. *Thematic Analysis: A Practical Guide* (1st ed.). SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP.

[5] John W. Creswell and Cheryl N. Poth. 2017. *Qualitative Inquiry and Research Design* (fourth ed.). SAGE Publications Inc.

[6] Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (April 2020), 7659–7666. doi:10.1609/aaai.v34i05.6267

[7] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How Culture Shapes What People Want From AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3613904.3642660

[8] Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2025. Do Generative AI Models Output Harm While Representing Non-Western Cultures: Evidence from a Community-Centered Approach. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24)*. AAAI Press, San Jose, California, USA, 476–489.

[9] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33, 2-3 (June 2010), 61–83. doi:10.1017/S0140525X0999152X

[10] Mingzhen Huang, Shan Jia, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. [n. d.]. Exposing Text-Image Inconsistency Using Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

[11] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K Reddy, and Sunipa Dev. 2024. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. *arXiv preprint arXiv:2401.06310* (2024).

[12] Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *arXiv preprint arXiv:2407.06863* (2024).

[13] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS 2024)*. https://doi.org/10.48550/arXiv.2404.16019

[14] Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating Cultural Differences into Large Language Models. doi:10.48550/arXiv.2402.10946 arXiv:2402.10946 [cs]

[15] Zachary Munn, Micah D. J. Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. Systematic Review or Scoping Review? Guidance for Authors When Choosing between a Systematic or Scoping Review Approach. *BMC Medical Research Methodology* 18, 1 (Nov. 2018), 143. doi:10.1186/s12874-018-0611-x

[16] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, 27730–27744.

[17] Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The Case for" Thick Evaluations" of Cultural Representation in AI. *arXiv preprint arXiv:2503.19075* (2025).

[18] Gerald Roche. 2019. Articulating Language Oppression: Colonialism, Coloniality and the Erasure of Tibet's Minority Languages. *Patterns of Prejudice* 53, 5 (Oct. 2019), 487–514. doi:10.1080/0031322X.2019.1662074

[19] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. doi:10.48550/arXiv.2406.05967 arXiv:2406.05967 [cs]

[20] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic Is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 160–171. doi:10.1145/3593013.3593985

[21] Lili Zhang, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024. Partiality and Misconception: Investigating Cultural Representativeness in Text-to-Image Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.