The Case for "Thick Evaluations" of Cultural Representation in AI

Rida Qadri, Mark Díaz, Ding Wang, Michael Madaio

Google Research
ridaqadri@google.com, markdiaz@google.com, drdw@google.com, madaiom@google.com

Abstract

Generative AI model outputs have been increasingly evaluated for their (in)ability to represent non-Western cultures. We argue that these evaluations often operate through reductive ideals of representation, abstracted from how people define their own representation and neglecting the inherently interpretive and contextual nature of cultural representation. In contrast to these 'thin' evaluations, we introduce the idea of *'thick evaluations'*: a more granular, situated, and discursive measurement framework for evaluating representations of social worlds in AI outputs, steeped in communities' own understandings of representation. We develop this evaluation framework through workshops in South Asia, by studying the 'thick' ways in which people interpret and assign meaning to AI-generated images of their own cultures. We introduce practices for thicker evaluations of representation that expand the understanding of representation underpinning AI evaluations and by co-constructing metrics with communities, bringing measurement in line with the experiences of communities on the ground.

Introduction

Generative AI (genAI) models, despite their popularity, have been shown to fail at inclusively representing different cultures in generated outputs—including images (Qadri et al. 2023; Mim et al. 2024; Mack et al. 2024) and text (Wang et al. 2024b; Pawar et al. 2024; Singh et al. 2024; Myung et al. 2025; Khanuja et al. 2021)—much like failures of representation in other AI systems (e.g., Katzman et al. 2023; Chien and Danks 2024; Wang et al. 2024a; Kay, Matuszek, and Munson 2015; Shelby et al. 2023; Crawford 2017).

To address these gaps, prior work has sought to evaluate the cultural representations in AI-generated output, but with few exceptions (e.g., Ghosh 2024; Qadri et al. 2023), mostly through quantified, metricized approaches to representation such as statistical similarities and benchmark-style scoring (Zhang et al. 2024; Li et al. 2024). The use of these methods presumes that representation is an objective construct with a quantifiable, definitive ground truth that outputs can be compared against (e.g., Kannen et al. 2024; Zhang et al. 2024) (for a critique of ground truth, see Muller et al. 2021). Given limitations of computational methods, representation is reduced to evaluations of basic recognition or factual generation of artifacts. When human feedback on representation is sought, it is often solicited through narrow, constrained, quantitative scales from crowdworkers who may not have the lived experiences to evaluate nuances of content's varied meanings to different social groups (Díaz et al. 2022).

However, this approach to measuring representation is in contravention to decades of scholarship in the social sciences that emphasizes the subjective nature of representation, where judgments about representation in visual media are constructed in conversation with the viewer's lived experiences and the broader context within which an image is viewed and published. (Hall 1989, 1997). In addition, the categories many AI researchers use for evaluations are often abstracted from the experiences of communities, constructed by AI researchers without engaging with communities to understand what axes of representation might be salient in their contexts. For instance, while many representational evaluations focus on skin tone, skin tone might not be a salient category of differentiation for social groups in many cultures. Thus, while existing approaches to evaluating representation in AI/ML might be useful for evaluating the accuracy of visual depictions of the physical world, they may not allow us to engage with communities' diverse desires for representations of their *social worlds*.

As generative image models increasingly are used to represent social worlds, what new evaluation approaches are necessary to meaningfully account for the diverse ways that people interpret and evaluate cultural representation in AI?

We argue that effectively evaluating cultural representation in AI images requires **'thick evaluations'**—a more granular, situated, discursively constructed approach to measurement, steeped in communities' own understanding of appropriate cultural representation. Our analysis draws on the 'thick vs thin' dichotomy introduced by philosopher Gilbert Ryle (1968), later taken on by Clifford Geertz (2008) and others (Ortner 1995; Love 2013; Jackson 2013; Riles 2000) to characterize descriptions of social worlds. Ryle's classic example of a wink illustrates this distinction: a 'thin description' merely describes the physical act of closing one eye by focusing on observable details, while a 'thick description' unpacks the social meaning of the act and its significance as a signal to interlocutors. Seen through

this framework, emerging methods for evaluating cultural representation in AI images involve 'thin evaluations'; i.e., suited to evaluating the observable aspects of the physical world contained in AI-generated images, but not necessarily the social signals embedded in those physical depictions.

To develop a 'thick evaluation' approach for cultural representation in AI images, we turned to the communities represented in images we seek to evaluate, recognizing their expertise and stakes in visual representations of their cultures. Given prior work on AI's failures of representation for South Asian cultures (e.g., Qadri et al. 2023; Ghosh et al. 2024), we conducted workshops with 37 participants in 3 South Asian countries to study the culturally-situated ways people assign meaning to, interpret, and evaluate the representation of their cultures in AI-generated images. We find that 1) people evaluate representation of social worlds not just through a singular category such as accuracy, but through multi-dimensional, fine-grained axes; 2) people's goals for evaluating cultural representation are situated in their social context, negotiated through dialogue with others and in response to broader societal discourse about their cultures; and 3) people use situated social knowledge of and experiences with social worlds to evaluate varying social meanings of images.

These findings demonstrate that thin evaluations alone cannot measure thick constructs like cultural representation, and they raise critical questions about the adequacy of existing evaluation paradigms for measuring cultural representation in AI images. As new methods and standards are being created for AI evaluations, this is a crucial moment to call for a thicker AI evaluation practice that reflexively interrogates epistemological underpinnings of AI evaluation practices, fundamentally rethinks whose expertise is included in the evaluation of AI systems, and opens up space for more interpretive qualitative evaluation methods. We provide pathways for such thickness by showing how AI researchers can interrogate the construct of cultural representation and the ways that that unobservable construct is operationalized in evaluation methods (cf. Jacobs and Wallach 2021). We provide empirical data on varied categories of representation people might evaluate images for, which shows the need for co-constructing evaluation methods with members of communities whose culture is being represented, to bring measurement in line with their experiences of AI images. We thus help AI practice move towards an ecosystem of both 'thick' and 'thin' evaluations, encouraging congruence between the construct being evaluated and the evaluation methods used. As AI technology is entering into the space of cultural production, our work suggests the need to develop new, thicker forms of evaluations of cultural representation, to better reflect how people consume and interpret images of their cultures.

Related Work

Evaluating Cultural Representation in AI

In order to understand the nature of AI models' failures of representation (e.g., Qadri et al. 2023; Mack et al. 2024), researchers have developed methods for evaluating representational failures (e.g., stereotyping) of AI systems (e.g., Katzman et al. 2023; Chien and Danks 2024; Wang et al. 2024a; Kay, Matuszek, and Munson 2015; Shelby et al. 2023; Crawford 2017; Harvey et al. 2024). Primarily, these evaluations have focused on language models (e.g., Blodgett et al. 2021; Hosseini, Palangi, and Awadallah 2023; Abbasi et al. 2019; Katzman et al. 2023; Chien and Danks 2024; Wang et al. 2024a; Wolfe et al. 2024), but they have also begun to include image models. For instance, emerging approaches for evaluating representation in images include developing quantified benchmark scores for goodness of representation (e.g., diversity (Kannen et al. 2024)), via methods like creating statistical similarity between generated images and reference images (Zhang et al. 2024; Ghosh 2024), calculating correlations in keywords (Li et al. 2024), measuring model frequency of generating stereotypical and offensive images of nationality groups (Jha et al. 2024), and calculating differences in frequency between the most and least common identities referenced in a set of model outputs (Lahoti et al. 2023). Human feedback has been sought via, for instance, using anonymized crowdworkers to score cultural biases in images (Basu, Babu, and Pruthi 2023; Ghosh, Lutz, and Caliskan 2024). For representation in images, these measures often focus on visual demographic diversity characteristics—for instance, Cho, Zala, and Bansal (2022) use automated skin tone and gender presentation classifiers to evaluate diversity of generated images.

However, this line of work follows a technosolutionist, positivist (Crabtree 2024) conception of representation; i.e., treating representation as something objective, stable across time and contexts, and quantifiable. For instance, using skin tone to measure diversity presumes skin tone is a useful proxy for race, and race is a universal category of differentiation, which may not be appropriate in all global cultures being represented. In contrast, researchers at the intersection of HCI and responsible AI have conducted qualitative evaluations of representation in genAI image models, identifying failures of representation through more participatory and community-centered methods, with evaluators who are from the same culture as the images they are evaluating (Qadri et al. 2023; Mim et al. 2024; Ghosh, Lutz, and Caliskan 2024). However, while these efforts have demonstrated the value of qualitative approaches for evaluating representation in generative image models, they do not examine what is meant by representation as a construct, nor do they critically interrogate what the participants' process of interpretive meaning-making suggests for methods for evaluating cultural representation in AI more broadly—both of which we focus on in this paper.

Recent research has shed new light on the complex process of measuring unobservable social constructs such as fairness, as well as the potential harms of reductive measurement approaches. For instance, Selbst et al. (2019) critique the definitions of fairness that lead to misleading abstractions in its evaluation. Recognizing the contested nature of the construct of fairness, Smith et al. (2024) argue for bringing in stakeholders' expertise for more meaningful and contextually grounded evaluations. Scholars like Jacobs and Wallach (2021) advocate for utilizing approaches to *measurement modeling* from the social sciences to bridge

the gap between unobservable constructs and their operationalization in evaluations of fairness in algorithmic systems (and for evaluation of generative AI (e.g., Wallach et al. 2025; Weidinger et al. 2025))—a framework we draw on in this paper through our discussion of the construct of cultural representation and how it is operationalized in evaluations.

Jacobs and Wallach (2021) argue that measurement modeling begins with a clear and robust articulation of the construct being evaluated—in their paper, fairness, and in our paper, cultural representation. However, in prior work evaluating cultural representation in text-to-image models (e.g., Zhang et al. 2024; Kannen et al. 2024; Jha et al. 2024; Lahoti et al. 2023; Basu, Babu, and Pruthi 2023), the construct of representation is often not interrogated or only implicitly defined. Although often left implicit, prior work operationalizes representation in various ways, as lack of bias (Wan et al. 2024), as geo-cultural similarity (Basu, Babu, and Pruthi 2023), and as imbalance (Su et al. 2009). More broadly than for cultural representation in text-to-image models, Chasalow and Levy (2021) argue that researchers often fail to make explicit how they conceptualize 'representative' or 'representation.' Such various ways of operationalizing the same construct may lead to evaluations that use the same term, but optimize for disparate goals (e.g., cultural awareness, cultural diversity, avoiding "cultural bias" (Li et al. 2024), etc), making it difficult to compare the effectiveness of different approaches.

Situated Conceptualizations of Representation

Decades of scholarship from the humanities have argued that representation in media, particularly visual media, cannot be defined through positivist approaches (e.g., Hall 1989, 1997; Barthes 1999). They argue that representation is not a static goal that can be objectively and quantitatively evaluated and achieved, but is instead an ongoing interpretive act whose meaning is contested and negotiated. In media studies, visual media, such as paintings, photography, or digital art, do not represent an objective reality but instead communicate multiple shades of meaning about the world (Hall 1997). That is, images have both denotative meanings (i.e., the literal subject of the image) as well as connotative meanings (i.e., the emotional subtext or symbolic meaning) the image was designed to convey—or the meaning it evokes in the viewers, which may be different than the intended meaning of the image (Hall 1997; Barthes 1999).

As a result, for media studies scholars, representation in visual media is a process of meaning-making, not a one-to-one depiction of the world as it is—and thus representation is a site for struggle *over* meaning (Hall 1989, 1997; Desai 2000). In this way, even a photograph of an event does not objectively convey a singular true meaning (Hall 1997), but instead conveys a "positional truth" (Abu-Lughod 1991) (or multiple such truths, given the interpretive nature of images), shaped by the creators' and viewers' cultural and historical context. Thus, visual understanding itself relies on a slippery relationship between words, images, and the concepts signified by those images (Barthes 1999).

As media studies scholars remind us, representation in public culture is a "zone of contestation" (Desai 2000), characterized by contests over which peoples and perspectives are made visible, how they are portrayed (and by whom), contests which are shaped by who has the power to control narratives of representation. Thus, in this paper, we explore how people evaluate representation in AI-generated images and what this representation means to them in their culture.

Thick vs Thin Modes of Understanding the Social

The 'thick/thin' dichotomy, originating in Gilbert Ryle's philosophical work (1968), has become a cornerstone of anthropological and social scientific inquiry. He argued that understanding human behavior requires moving beyond mere observation of physical details to interpretation, recognizing the "many-layered sandwich" of meaning embedded within even seemingly simple actions (Ryle 1968). This concept was further elaborated upon by Clifford Geertz (2008), who emphasized the importance of thick description in ethnographic research. Geertz argued that researchers must not only describe actions but also interpret them within their social or cultural context. A wink, for instance, could signify a conspiratorial gesture, a flirt, a parody, or a mere rehearsal, each carrying distinct meanings depending on the social context.

Sherry Ortner, in a chapter on "thick resistance," highlighted the limitations of thin approaches in capturing the complexities of social life, which require the *"richness, texture, and detail,"* of thickness (Ortner 1995, 1997). She argued that thin descriptions often overlook the *"internal politics of dominated groups, thin on the cultural richness of those groups, thin on the subjectivity—the intentions, desires, fears, projects—of the actors engaged in these dramas"* (Ortner 1995). Ortner advocated for thick descriptions that delve into the *"production, circulation, and consumption"* of cultural practices, recognizing their dynamic interplay with social and political realities. While the value of thickness in capturing the nuances of social phenomena has been widely acknowledged, scholars have also recognized the importance of thinness in discerning patterns and broader social structures. Annelise Riles Riles (2000), in her work on multi-sited ethnography, explored the challenges of achieving thick description when studying global phenomena dispersed across diverse cultures, calling attention to *"the limits of thickness as a disciplinary trope"* and to pay more attention to what the use of "thin composition" could yield. Anthropologist John Jackson similarly emphasized the value of thin description in providing a *"baseline empiricism"* and a starting point for social investigation (Jackson 2013). Thus, while thick and thin both have their place in social analysis, each one can only get you so far. In this paper, we draw on the thick/thin dichotomy to provide a framework for navigating the complexities of social inquiry when considering multiple approaches to evaluating cultural representation in AI images.

Methods

In this research, we treat the evaluation process as an object of study itself through community-centered, qualitative methods. These methodological choices respond to increasing calls for broader participation in AI development

and design (Delgado et al. 2023; Birhane et al. 2022a). We engaged with participants from three South Asian countries (Sri Lanka, Pakistan, and India) through a series of individual online evaluations and collective in-person workshops. These methods were chosen to allow us to move beyond binary evaluation metrics like thumbs up/down (cf. Collins et al. 2024); we instead sought to capture in-depth qualitative reflections and discussions that participants have as they interpret and evaluate AI-generated images. Given the discursive and collective nature of representation in media, we chose group workshops instead of individual interviews. To preserve participants' perspectives, we included an individual evaluation task prior to the workshops. Additionally, to ensure the annotation process reflected local cultural norms—including platform choice, framing of questions, and participant composition—we collaborated with local research partners in each region.

Participants and Sites

We chose our research sites for two main reasons. First, we aim to contribute to a growing body of scholarship that seeks to expand the evaluation of AI beyond Western-centric perspectives (Kak 2020; Mohamed, Png, and Isaac 2020; Sambasivan et al. 2021; Qadri et al. 2023; Ghosh et al. 2024). Second, we intend to move beyond the common "USA vs. India" dichotomy often found in comparative analyses including the Global South (e.g., Raman et al. 2008; Khairullah and Khairullah 2009). To this end, we focused on three South Asian countries: Sri Lanka, Pakistan, and India. They were selected due to the shared cultural histories, while also recognizing the vast diversity within South Asia.

We recruited participants through collaborations with local research partners deeply embedded in their communities. Our partners in Sri Lanka and Pakistan were university researchers with extensive research experience with marginalized local communities, while our partner in India was a research manager experienced in conducting workshops and focus groups. Recruitment began with emails to targeted lists, including local institutions, previous research participants that worked with those partners, and the partners' professional networks. Interested individuals completed a screening form detailing demographics and background.

Rather than attempting to exhaustively represent any single culture, our aim was to capture nuanced diversity and similarities about how people from different cultures interpret and assign meaning to their cultural representation. Thus our selection criteria took a purposive sampling approach (Blandford, Furniss, and Makri 2016), aiming to identify a diverse set of individuals with various intersecting identities relevant to the local context. We were not prescriptive regarding the exact numbers and form of diversity in our sample, but instead we relied on our partners to guide us towards axes of marginalization and social differentiation that might influence socio-cultural experiences in the country. These led to a sample that was highly contextualized in its diversity but not necessarily exhaustive of all forms of diversities. Minority and diversity here are interpreted in the context of the respective country. For instance, Hinduism is not a minority religion in India, but it is in Sri Lanka and Pakistan, while being Muslim is not a minority in Pakistan, but it is in Sri Lanka. 15 participants per country were selected. Ultimately, 11 participants each from Sri Lanka and Pakistan and 15 from India (spread across two workshops) took part. Participants' backgrounds are detailed in the supplementary material here **??**.

Study Activities

Our study comprised three parts: a pre-workshop **survey** to elicit prompts for culturally relevant images, a pre-workshop **evaluation task** to evaluate those images, and a **workshop** to discuss cultural representation in AI images.

Pre-workshop surveys. We asked each participant to create prompts for AI image generation, drawing on their unique cultural experiences by mixing salient identity markers (e.g., "Punjabi woman") with a regional landmark or cultural event, resulting in prompts such as "a group of Punjabi women in front of Wazir Khan Mosque in Lahore." This fostered diverse and culturally-relevant prompts and enabled participants to be experts in evaluating the resulting images.

Individual evaluation task. We wanted to give each participant space to individually evaluate the images about their cultures generated by their prompts, before joining the workshop. Thus, after we gathered the prompts from the participants, one of the authors generated images based on each participant's prompts and compiled a personal collection of prompt and images for each person. Participants were invited to reflect on each image and leave commentary on whether the image was a good representation of their cultural experience. We purposefully did not define "good" or "representation," since those were terms we wanted participants to define themselves during the workshops.

Workshop. To facilitate collective reflection and discussion on cultural representation in AI images, we convened four-hour workshops across three sites: two in-person workshops in Pakistan and Sri Lanka, and two online workshops in India where logistical constraints necessitated a virtual format. Online workshops were held via Google Meet.

Each workshop lasted roughly four hours and had three sections. First, participants engaged in a reflective discussion on their representational goals, drawing from the image-prompt pairs they had evaluated before the workshop. Next, they participated in prompting exercises, iteratively refining prompts through multi-turn interactions with a generative AI model capable of text-to-text, image-to-text, image-to-image, and text-to-image tasks. Finally, participants shared their experiences with image generation and representation in an evaluation discussion.

The pre-workshop evaluation task complemented the workshop by providing participants with space for individual reflection, to prepare for the group discussions. This combination allowed us to explore both individual interpretations and collective understandings of cultural representation, drawing on participants' role as experts in their cultural contexts. Individual evaluations captured participants' diverse perspectives on AI-generated cultural representation, while the workshops encouraged dialogue to uncover collective interpretations and contestations in meaning-making.

Data Analysis

All workshops were recorded and transcribed, and where necessary, translated by authors or local partners. All four authors participated in data preparation (e.g., data cleaning and transcribing) and analysis. We took a reflexive thematic analysis approach to analyze the workshop data, following Braun and Clarke (2006, 2021)—this includes the transcripts as well as participants' responses to the individual pre-workshop evaluation task. We met regularly as a group throughout the analysis period to inductively generate themes that captured patterns of shared meaning across workshop sections and workshops (Braun and Clarke 2021). We used the digital whiteboard Mural to iteratively cluster the codes into larger themes, discussing the relationship between codes and themes as we went and resolving any disagreements in synchronous group discussions. To attribute the quote we used in the following section to the participants, we have used anonymized participant identifiers: country code (IN for India, PK for Pakistan, and LK for Sri Lanka) concatenated with a number.

Findings

We first show the granular, multi-dimensional categories of cultural representation people draw on when evaluating images, complicating the singular constructs of representation in thin evaluations. We then explore how participants' goals for appropriate cultural representation were deeply situated within their specific cultural contexts, dynamically change over time, and were discursively constructed through collective dialogues. Finally, we examine the situated forms of social knowledge that underpinned these evaluations, emphasizing the importance of thick evaluations that draw on people's lived experiences and cultural expertise.

People evaluate cultural representation through multi-dimensional categories

When asked to evaluate cultural representativeness of AI-generated images, our participants drew on multiple dimensions of representation to evaluate, demonstrating the need for more fine-grained categories of cultural representation than singular constructs like accuracy of representation. We distill these into five dimensions, which point to the richer categories that thick evaluations could evaluate for: *incorrectness* (the accuracy of depicted physical objects—the closest to existing approaches (e.g., Zhang et al. 2024)), *missingness* (absence of iconic and expected cultural elements), *specificity* (whether the subject of the image was specific to their particular sub-culture), *coherence* (whether all elements of an image were appropriate given cultural and social norms), and *connotation* (the symbolic meanings and interpretations associated with an image). These expanded categories of cultural representation demonstrate the need for constructs that go beyond the physical world to capture the social worlds of a culture.

Incorrectness Participants evaluated the (in)correct representation of specific artifacts that have a singular corresponding physical existence outside the image, such as a building, a landmark, or a local landscape. This category has the most congruence with existing thin evaluations, as it seeks to evaluate depictions of artifacts with respect to their existence, and can often be answered with a binary or quantifiable metric since the 'ground truth' is more easily determinable than the following dimensions of representation. For instance, participants talked about how the depiction of the Lotus Tower in Sri Lanka can be incorrect or correct since there is only one Lotus Tower with a very particular form. As such, when attempting to generate images, Sri Lankan participants reflected on how they were not able to produce correct images of the Lotus Tower, a famous landmark and the tallest structure in Sri Lanka. However while incorrectness was relevant for evaluating objects with clear counterparts in the physical world, it was not used to assess other forms of cultural representation of the social world, indicating that evaluations of cultural representation encompass more than just factual correctness.

Missingness Similarly, participants evaluated whether images lacked the expected cultural elements that they felt contributed to a place's identity—what we refer to as missingness (see prior work on the related, though distinct, topic of "erasure": Qadri et al. 2023; Shelby et al. 2023; Katzman et al. 2023; Qadri et al. 2025). Participants felt that the absence of common features and structures in images meant to represent their culture meant that a given culture or location's essence or identity would not be communicated. LK-5, a participant from Sri Lanka, noted scenes that are commonplace in Sri Lanka that were curiously missing across a number of generated images: *"iconic places were never represented, like post offices and railway stations. Food items in Sri Lanka were never represented. Beaches were never reflected in any images, [but] we are an island."* In a similar vein, LK-11 pointed out that even nationally significant elements of the Sri Lankan landscape—such as visually distinct species of trees—were absent, but should be represented: *"so it should reflect more of our national items if we were to ask it to generate things about Sri Lanka, such as the Naa tree, Bo tree, national bird, animals."* This sense of missingness was closely tied to participants' ability to distinguish one place from another and capture its unique cultural character—i.e., its iconicity (cf. Barasch and Serrano 1992).

Specificity Participants evaluated the specificity of cultural representation—i.e., whether the depiction of cultural elements was specific to their contexts—especially for elements indicative of intersectional identities or subcultures, or cases where cultural artifacts were shared across cultures (but manifested in slightly different ways). For example, when evaluating images of women wearing *saris*, participants identified the regional nuances of *saris* that make them specific to their culture: *"I would say it's Marathi. A Marathi sari is a little different. So, the saris also have very different variations. Right. So these saris, particularly, make me feel like it's from the Northern region. Like that is more, I think, mainstream India"* (IN-3). Evaluating for cultural specificity meant that participants paid close attention to nuanced details that distinguished the cultural practices of their social worlds from others. For other participants, when evaluating cultural representation, it was important to point out

that a style of dress was more common in another part of the world than theirs: *"this has more Arab cultural influence, like wearing long dresses and wearing hijabs"* (PK-8). Similar to missingness, evaluations of specificity were crucial for accurately conveying the unique aspects of a particular culture, bringing with them layers of meaning and symbolism from their social worlds.

Coherence Participants also evaluated the coherence of the images of their culture—i.e., considering the extent to which various elements in an image were in alignment about which culture they were representing and in what ways. They evaluated (in)coherence in multiple ways, such as evaluating images for unrealistic combinations of cultural elements, misaligned behaviors with social and cultural norms, and anachronistic elements of their culture. Participants identified images that contained a mishmash of cultural signifiers that did not typically occur together, such as merging cultural details from different (sub)regions, subcultures, religions, or nationalities into one image: *"So the attire is of the people of Tamil. And the lanterns look Chinese, and [there is] no symbol of Sinhal[ese] new year"* (LK-10). Other participants in this workshop elaborated on the incoherence, pointing out that: *"the kids are wearing what normally Sri Lankan kids [are] wearing. But she is also having the [bindhi], and the elder person is kind of Sri Lankan, the mom is kind of European again. And the other kid is having totally different [clothing]—a sari on a male kid I think? So I kind of gave it a [score of] two out of five."* (LK-1)

(In)coherence was also about (mis)alignments of the perceived social behaviour in an image, where people were behaving or engaging in activities that did not align with participants' own experiences or perceptions of what was considered appropriate or logical to do in their cultures, given local social and cultural norms:

- *"I've never seen women wearing jewelry in [the] Press Club. Women go there for protest, [but they are] looking like they are here for a picnic."* [PK-8]
- *"Even the second picture, that is also wrong. That is shown like people are boating in that place. It is never done in a religious place like a gurudwara. This is never done."* [IN-5]
- *"Most people go to such tombs for tourism, but it seems there's a religious ceremony or speech or proselytizing happening here in these images [like] 'tableegh,' which does not happen in these tombs..."* [PK-9]

Another form of (in)coherence was temporal, when elements from different time periods were placed together, creating anachronisms. IN-5 noted that some elements of images were correct once, but no longer: *"No one covers their head [now]. My mother used to cover her head. My granny used to cover her head, but I have not seen any woman who's living in a village or in a rural area [do that] nowadays. It was a part of Punjabi culture, but not anymore."*

Connotations Finally, participants evaluated the connotations evoked by the images, recognizing that representation in visual media is not just depiction, but also a communicative act. While certain elements of images may be accurate in the sense of potentially occurring in the physical world, participants discussed how those images evoked connotations of the social world that they felt were not appropriate representations of their culture. In the workshop in Pakistan, participants noted that depictions of beards in Pakistani images evoked Western stereotypes about Pakistanis, and they raised concerns about promoting narrow ideas of Pakistani culture: *"Everyone has beards in these photos. Many people in Pakistan don't have a beard. Even in this [workshop] group... this brother does not have a beard, even my beard is so small"* (PK-11). Similarly, participants interpreted connotations of poverty from the representation of particular cultural artifacts, which were not inherent to the denotations of the visual images. For example, showing women in *saris* in India was associated with *"a typical traditional thinking about the Indian women since long ancient times"* (IN-4). In some evaluations, participants explicitly expressed frustration as to these images being a signal of 'them' seeing 'us' in a particular light: *"In none of the pictures we see women. I don't know what they think of us. Like we are from the 19th century and live 200 years ago"* (PK-11).

Goals for cultural representation are situated, dynamic, and negotiated

Goals for representation are situated and dynamic. The previous section showed the more granular categories needed to evaluate the construct of cultural representation in AI images. In this section, we identify how the goals for evaluating those constructs of cultural representation are developed through a dynamic conversation among the participants about their lived experiences and broader messaging about their culture. Participants emphasized that there was no 'objective' ideal for representation, but instead what constitutes meaningful cultural representation varies both across and within cultures and contexts. For instance, diversity of representation is one example of a representational goal generative models might aim for, but one participant noted how identities considered diverse in other contexts were, in fact, the dominant identities in her own: *"what might be diversity in the First World might actually be monotony in my area... what is diversity in a First World context or in the rest of India is the dominant culture where I come from"* (IN-15). For IN-15, when the model attempted to achieve a decontextualized form of diversity from a Western lens, it was inadvertently just depicting dominant cultures: *"In the First World, there are attempts at sort of including Muslim representations and all of that... But for example when I give a prompt as a Kashmiri woman, Muslims are the majority. Here [in Kashmir] diversity would look like something else... So here I don't want just women with hijabs represented. I also want Muslims without hijabs and non-Muslim women represented"* (IN-15). Thus, while representational diversity was important for participants, there was no one-size-fits all approach to diversity, as it needed to be contextualized within the social worlds of the users to be meaningful.

Representational goals dynamically shifted not just across countries but also within a country. For instance, IN10 from India stated the difficulty of evaluating whether the model had successfully represented a holiday, because people cele-

brated the holiday so differently in different regions: *"Even in the South, Ganesh chaturthi is celebrated, but the scale of it is very different. So if you look at Mumbai has very massive celebration during Ganesh chaturthi. I don't think (the celebration in) Bangalore is as big and I think some families would do a quiet celebration"* (IN-10). Participants also highlighted how even for the same person, visual representations of their own behavior would have to vary based on differences in the context they were in—such as, for instance, urban and rural, public and private, and variance in their individual adherence to cultural norms. To exemplify this, participant IN-3 mentioned how their attire would change if they were in rural India vs. urban India, saying *"I would be wearing this in my village, yes [but not in the city]"* (IN-3). Similarly, looking at a generated image of a woman in traditional Pakistani clothes, PK-7 noted that what a woman in Pakistan might wear in the bazaar would be different from what she wore elsewhere: *"We do see women around us wearing Western clothes, but in the bazaar you would see [women] dressed like this [in traditional clothes], so I don't expect women [in generated images] wearing a crop top, even though that is very me."*

Social worlds themselves, and thus their ideal representations, also dynamically change over time, constantly evolving in ways that are not as easy to evaluate as changes to the purely physical world. Participants noted that some aspects of cultural representation that models had generated could have been considered representative at one time, but were no longer the case, as the social world it was representing was evolving: *"[This] seems like Pakistan from the 70s. Pakistan has evolved. This is a very old Pakistan. We have a Western touch now also"* (PK-11). These issues with representation across time included attire that participants thought was outdated, architectural and building styles that felt like they were from another time, and modes of transport that did not exist anymore.

Goals for representation are constructed through discursive negotiation. Participants' goals for and understanding of cultural representation was actively negotiated in conversation with broader social narratives they had encountered in other visual media and through dialogue with other participants during the workshop. Even seemingly objective judgments of (in)correctness were contextualized by participants within discourses and messages they had encountered outside the specific image at hand. Some participants developed goals for representation in response to stereotypes about their cultures in the media—such as perceptions that South Asian cultures were undeveloped (e.g., PK-11), leading them to want generated images that showed more modernity. Or, when evaluating the failure of models to generate important landmarks from their cultures, participants were contextualizing these failures within perceived general power relations between the Global North and Global South that they had experienced.

Such representational goals were also reflected upon and negotiated in dialogue with other participants. One example of such negotiation was a debate that played out in one workshop on the tensions between combating stereotypes with what one participant referred to as 'very clean' positive depictions compared to what another participant termed as 'realism.' One participant asked their group about the desire for positive representations of their cityscapes: *"What do we expect out of AI? Do we want [the images] to wash away our sins? Not have electricity poles? Do we want [the images] to be very clean?"* (PK-4). This participant then went on to say they would not want a positive representation at all times because it would not be an accurate image. One participant echoed this conclusion, noting that *"there is a tension between creativity and artistic expression and realism. If you ask it to be too real, you are putting restrictions... on AI artistic capability"* (PK-2).

This tension between goals for representations to combat stereotypes and realism emerged concretely in images that were interpreted as depicting a less modern version of their culture. For instance, in generated images of Pakistani women, some participants noted that women wearing traditional clothes and not more 'Western clothes' would convey the impression that Pakistan was not a modern place. However, other participants felt this image still reflected some women, even if it did not represent everyone, and thus was important to retain. For instance, PK-7 argued *"You have to represent the culture as opposed to a small minority of women."* To which PK-11 responded, *"But in all pictures [of Pakistani women] we are seeing the same culture. One of the four images could be of a woman from a modern society."* Participants also acknowledged that there may not be consensus on what constituted desired cultural representation, but instead, people's goals for representation may be based on the communicative intentions of the group being asked. For instance, PK-1 noted that *"country officials or ambassadors... want the best images of Pakistan. They will want [images to] be whitewashed, but we want to have stray cats and [electricity] poles"* (PK-1).

While not every dialogue reached consensus, the act of dialogue helped shape their respective goals for representation. Participants noted the importance of creating spaces for discursive engagement in the evaluation of cultural representation, allowing for the emergence of diverse perspectives and the co-construction of meaning. As one participant noted, *"If you gave the same image to all five of us, we would be pointing out different elements of it... So, [in] a group discussion, you can do it more fruitfully than an individual even with a guideline. And when we move on to the next time, we will all have a more keen eye on it"* (LK-11). This suggests that approaches to evaluating cultural representation should engage with the discursively constructed and negotiated goals for what appropriate representation looks like.

People draw on situated knowledge and experiences to evaluate representation

In this section, we examine the specific types of knowledge participants drew upon for different evaluative categories outlined in the first section on multi-dimensional categories. As participants moved from evaluating more empirically demonstrable aspects of the physical world (e.g., is this the Lotus Tower?) to more culturally-situated judgments of the relationship between the physical and social worlds (e.g., is

this scene a positive representative of my city?), participants relied on diverse forms of knowledge and deep experience of their social worlds. For instance, evaluations of incorrectness required less cultural and social knowledge, becoming almost a form of pattern recognition. However, other forms of evaluation such as coherence or connotation required more specific knowledge, since they sought to evaluate not just thin concepts of physicality but thicker concepts of sociality. Understanding where social realities were (in)coherently displayed needed an understanding of appropriate behaviors and cultural norms that are often unstated (e.g., what lakes you would boat in and which you would not, or which generation of women might not be covering their head anymore). Even when visual elements of the physical world were being evaluated, they were often linked to social cues relying on hyper-specific visual elements: from the specific type of jewelry you would wear at particular celebrations to recognizing indigenous trees, local hairstyles, how people sit in a particular space, or even the birds that fly above a particular mosque and the potholes on a street.

Participants themselves explicitly underscored the need for cultural knowledge gained through lived experience within particular social worlds to be able to evaluate the representation of different elements of those social worlds. Participants noted that, given the rich and complex relationship between cultural artifacts and social worlds, it would be difficult for foreigners to understand the connotations or relationship between specific types of jewelry, clothing, food and cultural traditions like religious festivals. This knowledge was also often innate and implicit, and was not necessarily able to be made explicit in an evaluation rubric. When prompted to explain why a certain representation was wrong or problematic or unsatisfactory, people struggled to articulate exactly why what they were seeing was inaccurate or contextually inappropriate, saying expressions like, *"there's an energy"* (PK-7) or *"there's a vibe"* (IN-3, IN-14, LK-1). Participants also noted that tourists or other outsiders may have a preconceived notion of their countries that don't match the local reality. They pointed out that relying on people to evaluate cultural representation of images from cultures other than their own would lead to worse evaluations, as outsiders might fail to recognize harmful stereotypes embedded in images, instead mistaking them for accurate portrayals.

This recognition of outsiders' limitations also prompted participants to reflect on their own inability to judge the same axes of representation for other cultures. *"If someone asks me to evaluate Africa, I already have a bias[ed view] of Africa and I will evaluate accordingly"* (PK-11). While participants emphasized the importance of lived experience for accurate cultural representation, they also recognized the need for this experience to be granular and specific to intra-country social worlds. Even being from the same country, participants admitted that their personal experiences with other subcultures within their own country were often limited. Many shared examples of cultural norms and knowledge that they were not previously familiar with before this workshop. For instance, in the Pakistan workshop, participants discovered different ways of celebrating the holiday of Nauroz, surprised at their own discoveries in conversation with each other:

> PK-10: *In Skardu we celebrate it differently. We have contests with eggs. The one whose egg breaks loses. We have 10 day long celebrations.*
>
> PK-6: *We celebrate it very differently. We have people gathering, people playing music, playing sports. We have different instruments. Like tambourine or rubab.*
>
> PK-7: *I've never heard of Nauroz in Lahore or Karachi.*

These findings highlight the deeply situated cultural and social knowledge needed for evaluating appropriate cultural representation of social worlds in generated images.

Discussion

In this discussion section, we reflect on three implications of our findings for the practice of AI evaluations: (1) the need to reconsider the construct of representation underpinning AI evaluations by creating congruence between that construct and its measurement, (2) the importance of co-constructing methods for evaluating representation with communities, and (3) the tensions inherent in attempts to "thicken" evaluation practices for cultural representation in AI.

Thick Methods for Thick Constructs of Cultural Representation

Our findings highlight the urgent need for a critical reappraisal of how we conceptualize and operationalize representation in AI evaluations. Instead of representation being a singular fixed concept or objective truth, participants interpreted it in situated and negotiated ways via multiple granular categories. Our participants empirically demonstrated that what gets bundled into the 'suitcase word' of representation (cf. Chasalow and Levy 2021) are multiple unobservable constructs of cultural representation, each requiring different approaches to measurement and metrics. Our findings thus suggest that AI evaluations typically conflate two constructs of representation—one that is thinner, which focuses on empirical accuracy of physical worlds, and one that is thicker, such as "coherence," which reflects the internal social consistency and plausibility of the social world depicted in an image.

While thin evaluations may be suitable for assessing "thinner" constructs like the factual incorrectness of images' correspondence with an empirical reality or ground truth (e.g., landmarks), thicker constructs, such as those related to cultural norms, social relations, and power dynamics, necessitate thick evaluations that can capture the deeper layers of social meaning within images. We saw in our findings that as the representational category became "thicker" (e.g., moving from "incorrectness" to "connotation"), the evaluation process became more subjective and reliant on shared cultural understanding. Yet the field of AI is increasingly using thin metrics to evaluate thick concepts—i.e., using metrics and benchmarks that are better suited for evaluations of the physical world to instead evaluate the representation of social worlds. We thus advocate for an

evaluation ecosystem that creates congruence between the construct (i.e., cultural representation, or its dimensions) and the measurement method, which complements thin evaluations with thick evaluations. In the words of Jackson (2013), *"[T]hin description is the necessary starting point for social investigation but not nearly enough all by itself,"* or, as Ryle (1968) argues, *"thick description is a many layered sandwich, of which only the bottom slice is catered for by the thinnest description."*

Co-Constructing Evaluations of AI's Cultural Representation with Communities

Our study shows that there is no objective conceptualization of representation to optimize for in images. What should ideally appear when a model is prompted for "Indian," "Pakistani," or "American" is a culturally- and contextually-contingent choice. Thus, most representational evaluations require evaluators to make interpretations, involving culturally-situated judgment, rooted in social experience. Who gets to decide what representation is and how a culture should be depicted? Who has the power to shape meaning in the struggle over meaning-making that is representation in media (cf. Desai 2000; Hall 1997; Abu-Lughod 1991)? Currently, the power of this choice lies with AI researchers and anonymous evaluators who often have limited experience with the social worlds subject to evaluation. We argue, as with AI fairness more broadly, that we must *"bring the people back in"* to AI measurement of representation (Denton et al. 2020; Smith et al. 2024), to make it more contextual and remove translation gaps between abstract metrics and the actual risks and harms they seek to measure (Jacobs and Wallach 2021; D'ignazio and Klein 2023).

Although some prior work has conducted qualitative evaluations of representation in AI, they have primarily focused on on differences in stakeholders' judgments or decisions compared to some gold standard, with less focus on differences in the underlying *evaluative processes* that inform those judgments. As an example, disagreement itself is a robust area of research in NLP and AI, with much focus on the sociocultural influences that shape human judgment (Davani, Díaz, and Prabhakaran 2022; Díaz et al. 2022). Studying the culturally-situated ways that people evaluate and understand representation in our study allowed us to suggest metrics and measures of representation that are more in line with the ways people who actually experience and consume images interpret representation of their culture. Thus, our findings echo researchers like Smith et al. (2024) who call for AI measurement to come in line with the actual experiences of people on the ground—not just the experiences of researchers—and integrate their experiences, knowledge, and expertise (cf. Díaz and Smith 2024) in the measurement process. The aim is not simply to apply the same metrics in different ways, but to co-construct metrics and measures with stakeholders. Integration of expertise for representation must go beyond large-scale, globally-deployed surveys often used by AI researchers. This approach offers only a very limited mechanism to integrate cultural knowledge from particular communities into evaluative frameworks, because the metrics themselves are developed by AI researchers. Annotation evaluation pipelines are structured to align resulting data with what data requesters define to be ground truth (Miceli, Schuessler, and Yang 2020; Posada 2023; Chandhiramowuli et al. 2024; Muller et al. 2021). As a result, annotators or evaluators must constrain their assessments within the feedback categories set by researchers. Evaluations outside of these constraints for evaluation tasks are conflated with other concerns, or dismissed altogether (Posada 2023).

Building on work on subjectivity in annotation (Cabitza, Campagner, and Basile 2023; Davani, Díaz, and Prabhakaran 2022) and work in HCI that contests the idea of singular ground truth (Muller et al. 2021; Gordon et al. 2022), our study suggests opportunities for developing evaluation methods for representation that leverage discursiveness and deliberation (cf. Shen et al. 2022). Bergman et al. (2024) have highlighted the value of discursive deliberation in eliciting individuals' justifications for the views they espouse. To do this, evaluative methods might actively encourage collective conversation and disagreement among evaluators through interpretive methods like workshops in addition to thin methods like benchmarks. In our study, discursive evaluation helped to fill knowledge gaps, such as whether people in different regions celebrated certain occasions in the same way. However, our discursive evaluations also revealed and helped to resolve disagreements that were ontological in nature, about the very nature of representation. Ontological disagreement here was not simply a matter of knowledge gaps, but instead fundamental differences in understanding what an image represents and what it *should* represent. Discursive evaluation processes both shed light on ontological disagreements and shift understandings in real time. Thus, dialogue and discussion can change or reaffirm evaluative judgments evaluators might have held before—changing the nature of feedback evaluators might give in individual annotation tasks.

Making Space for Thickness in AI Practice

As we build the case for thick evaluations for cultural representations, we also reflect on the epistemological hurdles that may emerge if thick evaluations were to be adopted by AI developers. For one, thick evaluations are at odds with incentives and values like generalization and scale in computing (Birhane et al. 2022b; Hanna et al. 2020). Thick evaluations demand a deep engagement with situated cultural contexts (cf. Arzberger et al. 2024), requiring time and investment, and acknowledging the diversity of interpretations and meanings associated with representation. They embrace subjectivity and multiple interpretations, acknowledging that there is no single "correct" way to represent a social world. This all limits their scalablity and generalizability—but at the same time, they also empirically *demonstrate* the limits of scale and generalizability. As Tsing (2012) reminds us, the multiplicity of social worlds resists attempts at hegemonic structuring into scalable units.

For representational evaluations to become 'thicker,' AI practitioners and researchers will also have to stop drawing on the epistemic cultures of positivism and technosolutionism prevalent in ML and computing (Morozov 2013; Lindtner, Bardzell, and Bardzell 2016; Cunning-

ham et al. 2023). These epistemologies in AI have been extensively critiqued in AI fairness. For instance, annotation practices rarely account for social histories of race and gender (Scheuerman et al. 2020), and more broadly, algorithms fail to capture social nuances and lived realities (e.g., Broussard 2018; Mattern 2021; Arzberger et al. 2024). The critiques of thin-ness our findings proffer are also in line with critiques of positivism in the social sciences. For instance, Babones (2016) argues against treating humans and society as a *"knowable objective reality, represented reasonably well by the variables"* and Kitchin (2014) argues that positivism produces social analysis that *"is reductionist, functionalist and ignores the effects of culture, politics, policy, governance and capital."* In this paper, we demonstrate how current approaches to evaluating representation thus ignore extensive critiques of quantitative measurement in the social sciences, which caution against attempting to formalize fundamentally social processes, like cultural representation, into quantifiable metrics or objective functions that assume a singular empirical reality (Steinmetz 2005). Our findings demonstrate how limited the prevailing paradigms of objectivity and quantifiability are when attempting to evaluate social worlds, and how much richer AI evaluations could be if we embraced the pluralism of thick evaluations, with their emphasis on situated context, subjectivity, and lived experiences. Such a move would require de-centering objectivity, acknowledging the situatedness of evaluations of cultural representation.

Our findings also show the need for a fundamental shift in how we understand and value knowledge in AI practice. The practice of thickness needs communities' participation in co-constructing evaluations. That is why in this paper we draw on interpretive, dialogue-based methods to capture the discursive processes involved in negotiations of the meaning of images. Our methods also recognized the expertise brought by local experts and stakeholders who could partner with us to create a sample situated in the diverse social worlds of participants. However, adopting this practice would require AI researchers to recognize and embrace pluralistic ways of knowing and understanding the world. Integrating qualitative, situated knowledge into the predominantly quantitative realm of AI necessitates a shift in mental models and a confrontation with deeply ingrained assumptions about considerations of expertise (cf. Madaio et al. 2024; Díaz and Smith 2024). We take inspiration from those like Agre who call for a critical technical practice around computing (Agre 2014; Malik and Malik 2022), to call for a practice of thickness not just in measurement but also in AI writ large. This would require questioning the singular focus on scale and generalization, embracing qualitative methods and epistemologies, actively co-creating AI with communities and reflexively interrogating the epistemological values embedded in our evaluation practices.

Limitations and Future Work

This work has several limitations, which future work may address. First, our participants were from three countries in South Asia; bringing their own cultural and individual perspectives on the most salient elements for which to evaluate representation. Since generalizability is not the goal of interpretive qualitative research, which instead aims for *transferability*, (Soden, Toombs, and Thomas 2024; Drisko 2025), we do not make any claims of generalizability of this research to other locales. Thus, future research could help shed light on the extent to which similar taxonomies of thick evaluations may be transferrable (Drisko 2025) or extendable for representational evaluations with different contexts or populations. In addition, we focused on text-to-image AI systems in this study. Although we believe our work could speak to thick evaluations of representation in other AI modalities (e.g., text, video, audio, or multi-modal systems), it is an open question of precisely how such thick evaluations in other modalities ought to be conducted, or how the approach ought to change, if so. In this work, we discuss how certain elements of cultural representation may be best evaluated by thick evaluations, while others (e.g., the "incorrectness" dimension) may be better suited to thin evaluations, or correspondence to a ground truth; however, future work should be conducted to validate this hypothesis, to better understand which elements of cultural representation are best evaluated with which approaches. Finally, although in this paper, we map out a space for conducting thick evaluations, future work should explore how best to incorporate the results of these evaluations into shaping the design and development of generative AI systems, in order to avoid the failures of cultural representation that motivated this work.

Conclusion

As AI technologies are being proposed to engage in cultural production, and research continues to highlight the failures of AI to adequately represent all cultures, our work suggests the need to develop new, thicker forms of evaluations of cultural representation, to better reflect how people consume and interpret images of their cultures. Thus, in this paper, we introduce a framework for thick evaluations of cultural representation, developed in conversation with participants in South Asia, that can move the field beyond existing reductive ideals of representation. Showing that representation is a process of meaning-making through the interaction between image, viewer, and context, we argue that measurement of representation in AI needs to encompasses more than just factual correctness, creating practices that bridge the disconnect between constructing representation as a technical goal and understanding it as a social goal. Developing congruence between the thickness of representation as a concept and thickness of the evaluation method will ensure AI practitioners avoid creating brittle, impoverished evaluations and mitigations for representation in the design of AI systems, and, ideally, lead towards AI systems that are better able to represent the plurality of social worlds of all cultures.

References

Abbasi, M.; Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2019. Fairness in representation: quantifying stereotyping as a representational harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, 801–809. SIAM.

Abu-Lughod, J. L. 1991. *Before European hegemony: the world system AD 1250-1350*. Oxford University Press, USA.

Agre, P. E. 2014. Toward a critical technical practice: Lessons learned in trying to reform AI. In *Social science, technical systems, and cooperative work*, 131–157. Psychology Press.

Arzberger, A.; Buijsman, S.; Lupetti, M. L.; Bozzon, A.; and Yang, J. 2024. Nothing Comes Without Its World–Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 61–73.

Babones, S. 2016. Interpretive quantitative methods for the social sciences. *Sociology*, 50(3): 453–469.

Barasch, M.; and Serrano, L. 1992. *Icon: Studies in the History of an Idea*. NYU Press.

Barthes, R. 1999. Rhetoric of the Image. *Visual culture: The reader*, 33–40.

Basu, A.; Babu, R. V.; and Pruthi, D. 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.

Bergman, S.; Marchal, N.; Mellor, J.; Mohamed, S.; Gabriel, I.; and Isaac, W. 2024. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports*, 14(1): 6616.

Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022a. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022b. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.

Blandford, A.; Furniss, D.; and Makri, S. 2016. *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Braun, V.; and Clarke, V. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology*, 18(3): 328–352.

Broussard, M. 2018. *Artificial unintelligence: How computers misunderstand the world*. mit Press.

Cabitza, F.; Campagner, A.; and Basile, V. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6860–6868.

Chandhiramowuli, S.; Taylor, A. S.; Heitlinger, S.; and Wang, D. 2024. Making Data Work Count. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–26.

Chasalow, K.; and Levy, K. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 77–89.

Chien, J.; and Danks, D. 2024. Beyond Behaviorist Representational Harms: A Plan for Measurement and Mitigation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 933–946.

Cho, J.; Zala, A.; and Bansal, M. 2022. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers. *CoRR*, abs/2202.04053.

Collins, K. M.; Kim, N.; Bitton, Y.; Rieser, V.; Omidshafiei, S.; Hu, Y.; Chen, S.; Dutta, S.; Chang, M.; Lee, K.; et al. 2024. Beyond Thumbs Up/Down: Untangling Challenges of Fine-Grained Feedback for Text-to-Image Generation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 293–303.

Crabtree, A. 2024. H is for Human and How (Not) To Evaluate Qualitative Research in HCI. *arXiv preprint arXiv:2409.01302*.

Crawford, K. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Cunningham, J.; Benabdallah, G.; Rosner, D.; and Taylor, A. 2023. On the Grounds of Solutionism: Ontologies of Blackness and HCI. *ACM Trans. Comput.-Hum. Interact.*, 30(2).

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23.

Denton, E.; Hanna, A.; Amironesei, R.; Smart, A.; Nicole, H.; and Scheuerman, M. K. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*.

Desai, D. 2000. Imaging difference: The politics of representation in multicultural art education. *Studies in Art Education*, 41(2): 114–129.

Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2342–2351.

Díaz, M.; and Smith, A. D. 2024. What Makes An Expert? Reviewing How ML Researchers Define" Expert". In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 358–370.

D'ignazio, C.; and Klein, L. F. 2023. *Data feminism*. MIT press.

Drisko, J. W. 2025. Transferability and generalization in qualitative research.

Geertz, C. 2008. Thick description: Toward an interpretive theory of culture. In *The cultural geography reader*, 41–51. Routledge.

Ghosh, S. 2024. Interpretations, Representations, and Stereotypes of Caste within Text-to-Image Generators. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 490–502.

Ghosh, S.; Lutz, N.; and Caliskan, A. 2024. "I Don't See Myself Represented Here at All": User Experiences of Stable Diffusion Outputs Containing Representational Harms across Gender Identities and Nationalities. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 463–475.

Ghosh, S.; Venkit, P. N.; Gautam, S.; Wilson, S.; and Caliskan, A. 2024. Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 476–489.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

Hall, S. 1989. Cultural identity and cinematic representation. *Framework: The Journal of Cinema and Media*, (36): 68–81.

Hall, S. 1997. *Representation: Cultural Representations and Signifying Practices*. London: SAGE Publications.

Hanna, A.; Denton, E.; Smart, A.; and Smith-Loud, J. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 501–512.

Harvey, E.; Sheng, E.; Blodgett, S. L.; Chouldechova, A.; Garcia-Gathright, J.; Olteanu, A.; and Wallach, H. 2024. Gaps Between Research and Practice When Measuring Representational Harms Caused by LLM-Based Systems. *arXiv preprint arXiv:2411.15662*.

Hosseini, S.; Palangi, H.; and Awadallah, A. H. 2023. An empirical study of metrics to measure representational harms in pre-trained language models. *arXiv preprint arXiv:2301.09211*.

Jackson, J. L. 2013. *Thin description: ethnography and the African Hebrew Israelites of Jerusalem*. Harvard University Press.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 375–385.

Jha, A.; Prabhakaran, V.; Denton, R.; Laszlo, S.; Dave, S.; Qadri, R.; Reddy, C.; and Dev, S. 2024. ViSAGe: A Global-Scale Analysis of Visual Stereotypes in Text-to-Image Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12333–12347.

Kak, A. 2020. " The Global South is everywhere, but also always somewhere" National Policy Narratives and AI Justice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 307–312.

Kannen, N.; Ahmad, A.; Andreetto, M.; Prabhakaran, V.; Prabhu, U.; Dieng, A. B.; Bhattacharyya, P.; and Dave, S. 2024. Beyond Aesthetics: Cultural Competence in Text-to-Image Models. *arXiv preprint arXiv:2407.06863*.

Katzman, J.; Wang, A.; Scheuerman, M.; Blodgett, S. L.; Laird, K.; Wallach, H.; and Barocas, S. 2023. Taxonomizing and measuring representational harms: A look at image tagging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14277–14285.

Kay, M.; Matuszek, C.; and Munson, S. A. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, 3819–3828.

Khairullah, D. H.; and Khairullah, Z. Y. 2009. Cross-cultural analysis of gender roles: Indian and US advertisements. *Asia Pacific Journal of Marketing and Logistics*, 21(1): 58–75.

Khanuja, S.; Bansal, D.; Mehtani, S.; Khosla, S.; Dey, A.; Gopalan, B.; Margam, D. K.; Aggarwal, P.; Nagipogu, R. T.; Dave, S.; Gupta, S.; Gali, S. C. B.; Subramanian, V.; and Talukdar, P. 2021. MuRIL: Multilingual Representations for Indian Languages. arXiv:2103.10730.

Kitchin, R. 2014. Big Data, new epistemologies and paradigm shifts. *Big Data Society*, 1(1): 1–12.

Lahoti, P.; Blumm, N.; Ma, X.; Kotikalapudi, R.; Potluri, S.; Tan, Q.; Srinivasan, H.; Packer, B.; Beirami, A.; Beutel, A.; et al. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. *arXiv preprint arXiv:2310.16523*.

Li, H.; Jiang, L.; Dziri, N.; Ren, X.; and Choi, Y. 2024. CULTURE-GEN: Revealing Global Cultural Perception in Language Models through Natural Language Prompting. *arXiv preprint arXiv:2404.10199*.

Lindtner, S.; Bardzell, S.; and Bardzell, J. 2016. Reconstituting the utopian vision of making: HCI after technosolutionism. In *Proceedings of the 2016 chi conference on human factors in computing systems*, 1390–1402.

Love, H. 2013. Close reading and thin description. *Public culture*, 25(3): 401–434.

Mack, K. A.; Qadri, R.; Denton, R.; Kane, S. K.; and Bennett, C. L. 2024. "They only care to show us the wheelchair": disability representation in text-to-image AI models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–23.

Madaio, M.; Kapania, S.; Qadri, R.; Wang, D.; Zaldivar, A.; Denton, R.; and Wilcox, L. 2024. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1544–1558.

Malik, M.; and Malik, M. M. 2022. Critical technical awakenings. *Journal of Social Computing*, 2(4): 365–384.

Mattern, S. 2021. *A City Is Not a Computer: Other Urban Intelligences*. Princeton University Press.

Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–25.

Mim, N. J.; Nandi, D.; Khan, S. S.; Dey, A.; and Ahmed, S. I. 2024. In-Between Visuals and Visible: The Impacts of Text-to-Image Generative AI Tools on Digital Image-making Practices in the Global South. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33: 659–684.

Morozov, E. 2013. *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

Muller, M.; Wolf, C. T.; Andres, J.; Desmond, M.; Joshi, N. N.; Ashktorab, Z.; Sharma, A.; Brimijoin, K.; Pan, Q.; Duesterwald, E.; et al. 2021. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–16.

Myung, J.; Lee, N.; Zhou, Y.; Jin, J.; Putri, R. A.; Antypas, D.; Borkakoty, H.; Kim, E.; Perez-Almendros, C.; Ayele, A. A.; Gutiérrez-Basulto, V.; Ibáñez-García, Y.; Lee, H.; Muhammad, S. H.; Park, K.; Rzayev, A. S.; White, N.; Yimam, S. M.; Pilehvar, M. T.; Ousidhoum, N.; Camacho-Collados, J.; and Oh, A. 2025. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. arXiv:2406.09948.

Ortner, S. B. 1995. Resistance and the problem of ethnographic refusal. *Comparative studies in society and history*, 37(1): 173–193.

Ortner, S. B. 1997. Thick resistance: Death and the cultural construction of agency in Himalayan mountaineering. *Representations*, 59: 0.

Pawar, S.; Park, J.; Jin, J.; Arora, A.; Myung, J.; Yadav, S.; Haznitrama, F. G.; Song, I.; Oh, A.; and Augenstein, I. 2024. Survey of Cultural Awareness in Language Models: Text and Beyond. arXiv:2411.00860.

Posada, J. 2023. Platform Authority and Data Quality: Who Decides What Counts in Data Production for Artificial Intelligence. Technical report, Technical Report. Berggruen Institute and Global Affairs Canada.

Qadri, R.; Davani, A. M.; Robinson, K.; and Prabhakaran, V. 2025. Risks of Cultural Erasure in Large Language Models. *arXiv preprint arXiv:2501.01056*.

Qadri, R.; Shelby, R.; Bennett, C. L.; and Denton, E. 2023. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 506–517. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Raman, P.; Harwood, J.; Weis, D.; Anderson, J. L.; and Miller, G. 2008. Portrayals of older adults in US and Indian magazine advertisements: A cross-cultural comparison. *The Howard Journal of Communications*, 19(3): 221–240.

Riles, A. 2000. *The Network Inside Out*. University of Michigan Press.

Ryle, G. 1968. *The thinking of thoughts*. 18. [Saskatoon]: University of Saskatchewan.

Sambasivan, N.; Arnesen, E.; Hutchinson, B.; Doshi, T.; and Prabhakaran, V. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 315–328.

Scheuerman, M. K.; Wade, K.; Lustig, C.; and Brubaker, J. R. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW1): 1–35.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.

Shen, H.; Wang, L.; Deng, W. H.; Brusse, C.; Velgersdijk, R.; and Zhu, H. 2022. The model card authoring toolkit: Toward community-centered, deliberation-driven AI design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 440–451.

Singh, S.; Romanou, A.; Fourrier, C.; Adelani, D. I.; Ngui, J. G.; Vila-Suero, D.; Limkonchotiwat, P.; Marchisio, K.; Leong, W. Q.; Susanto, Y.; Ng, R.; Longpre, S.; Ko, W.-Y.; Smith, M.; Bosselut, A.; Oh, A.; Martins, A. F. T.; Choshen, L.; Ippolito, D.; Ferrante, E.; Fadaee, M.; Ermis, B.; and Hooker, S. 2024. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. arXiv:2412.03304.

Smith, J. J.; Satwani, A.; Burke, R.; and Fiesler, C. 2024. Recommend Me? Designing Fairness Metrics with Providers. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2389–2399. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Soden, R.; Toombs, A.; and Thomas, M. 2024. Evaluating interpretive research in HCI. *Interactions*, 31(1): 38–42.

Steinmetz, G. 2005. Introduction. Positivism and its others in the social sciences.

Su, P.; Mao, W.; Zeng, D.; Li, X.; and Wang, F.-Y. 2009. Handling class imbalance problem in cultural modeling. In *2009 IEEE international conference on intelligence and security informatics*, 251–256. IEEE.

Tsing, A. L. 2012. On nonscalability: The living world is not amenable to precision-nested scales. *Common knowledge*, 18(3): 505–524.

Wallach, H.; Desai, M.; Cooper, A. F.; Wang, A.; Atalla, C.; Barocas, S.; Blodgett, S. L.; Chouldechova, A.; Corvi, E.; Dow, P. A.; et al. 2025. Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge. *arXiv preprint arXiv:2502.00561*.

Wan, Y.; Subramonian, A.; Ovalle, A.; Lin, Z.; Suvarna, A.; Chance, C.; Bansal, H.; Pattichis, R.; and Chang, K.-W. 2024. Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation. *arXiv preprint arXiv:2404.01030*.

Wang, A.; Bai, X.; Barocas, S.; and Blodgett, S. L. 2024a. Measuring machine learning harms from stereotypes: requires understanding who is being harmed by which errors in what ways. *arXiv preprint arXiv:2402.04420*.

Wang, W.; Jiao, W.; Huang, J.; Dai, R.; Huang, J.-t.; Tu, Z.; and Lyu, M. 2024b. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6349–6384. Bangkok, Thailand: Association for Computational Linguistics.

Weidinger, L.; Raji, I. D.; Wallach, H.; Mitchell, M.; Wang, A.; Salaudeen, O.; Bommasani, R.; Ganguli, D.; Koyejo, S.; and Isaac, W. 2025. Toward an evaluation science for generative AI systems. *arXiv preprint arXiv:2503.05336*.

Wolfe, R.; Dangol, A.; Howe, B.; and Hiniker, A. 2024. Representation Bias of Adolescents in AI: A Bilingual, Bicultural Study. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1621–1634.

Zhang, L.; Liao, X.; Yang, Z.; Gao, B.; Wang, C.; Yang, Q.; and Li, D. 2024. Partiality and Misconception: Investigating Cultural Representativeness in Text-to-Image Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–25.