AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia

Rida Qadri
Google Research
San Francisco, California, USA

Renee Shelby
Google Research
San Francisco, California, USA

Cynthia L. Bennett
Google Research
New York, New York, USA

Emily Denton
Google Research
New York, New York, USA

ABSTRACT

This paper presents a community-centered study of cultural limitations of text-to-image (T2I) models in the South Asian context. We theorize these failures using scholarship on dominant media regimes of representations and locate them within participants' reporting of their existing social marginalizations. We thus show how generative AI can reproduce an outsiders gaze for viewing South Asian cultures, shaped by global and regional power inequities. By centering communities as experts and soliciting their perspectives on T2I limitations, our study adds rich nuance into existing evaluative frameworks and deepens our understanding of the culturally-specific ways AI technologies can fail in non-Western and Global South settings. We distill lessons for responsible development of T2I models, recommending concrete pathways forward that can allow for recognition of structural inequalities.

KEYWORDS

human-centered AI, AI harms, cultural harms of AI, text-to-image models, generative AI, non-western AI fairness, South Asia, qualitative research in AI, failure modes

ACM Reference Format:
Rida Qadri, Renee Shelby, Cynthia L. Bennett, and Emily Denton. 2023. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia . In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3593013.3594016

1 INTRODUCTION

Emerging FAccT scholarship points to the need for reevaluating the field's dominant methods and frameworks for understanding and evaluating AI harms. For instance, there are growing calls for more community-centered work [62] and a re-orientation towards non-Western frameworks of fairness [4, 60, 76, 84, 99, 99, 100]. However, empirical studies collaboratively investigating AI harms with diverse, global communities are less common, continuing the disconnect between dominant evaluation approaches and the lived experiences of impacted communities [19].

In response, we conducted a community-centered study of cultural limitations of text-to-image (T2I) models in the South Asian context, with 36 participants from Pakistan (n = 15), India (n = 13), and Bangladesh (n = 8). Through two-part focus groups, participants co-designed T2I prompts and collectively reflected on model outputs. This study design offered participants agency to articulate their own understandings of model limitations, failures, and potential impacts, drawing from their local cultural knowledge and situated experiences with South Asian representations. Participant conversation and reflections foreground three broad failure modes: failing to generate cultural subjects, amplifying hegemonic cultural defaults, and perpetuating cultural tropes. We contextualize these failures within literature on media and cultural studies, that study the power of media's *regimes of representations* [48, p. 234]: controlling narratives and discourses about particular social groups. Our study shows how cultural limitations of T2I models can participate in and scale such existing harmful media regimes of representation and amplify experiences of socio-cultural marginalization. While the T2I failure modes foregrounded by participants are not necessarily unique to South Asia, their specific articulation in South Asia is contextual and socially situated in the global and regional power dynamics that shape the region.

It is important to note, however, this study is not a systematic evaluation of T2I model capabilities. As such, we do not position our findings as definitive commentary on any one or all T2I models. Nonetheless, our study, albeit a limited form of community engagement, encourages important methodological reflection on common machine learning evaluative and benchmarking practices. Emerging empirical work on T2I evaluations focuses on quantitative evaluation [28, 45], often with bias metrics pre-determined by researchers and practitioners [14]. While such approaches enable evaluating models at "scale," without community input they risk centering un-nuanced notions of harm that do not fully account for on-the-ground community experiences in different cultural contexts.

Our work contributes to multiple strands of responsible AI research. First, we respond to prior calls to re-orient algorithmic fairness outside Western contexts [101, 102] by offering the first empirical study of T2I performance in South Asia. Second, by soliciting community perspectives on T2I limitations, we identify novel T2I failure modes and connect model limitations to communities' lived experiences, expanding the field's understanding of cultural

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594016

harms. Third, by historicizing emerging generative image technologies within scholarship on the politics of representation, we identify particular regimes of representation within T2I models, revealing how these regimes draw from and perpetuate marginalizing discourses. Thus, our study offers an example of how qualitative, community-centered research strengthens responsible AI practice through centering local knowledge and expertise.

2 THEORETICAL BACKGROUND

2.1 Text-to-Image Models and AI Harms

Text-to-image (T2I) generative models allow users to create photo-realistic images from free-form and open-ended text prompts [87, 89, 96, 121], typically relying on web-scale datasets. Such datasets have been shown to reflect social stereotypes, inequalities, and hierarchies [17, 18, 83], raising concerns about T2I models similarly fostering representational and cultural harms [105, 111, 114]. However, unlike other generative AI, such as language [10, 117] or image caption models [116], computing researchers have yet to articulate the broader landscape of potential harms for generative image models. While empirical research on T2I models is still nascent, studies suggest they can reinforce social hierarchies and replicate dominant stereotypes along axes of gender, skin tone, and culture [7, 14, 28, 119]. Our work complements and extends this work, by offering the first empirical study of T2I models that centers and engages non-Western communities.

2.2 Non-Western and Community-Centered Fairness

There is a growing body of scholarship calling attention to the dominance of Western perspectives and experiences embedded within responsible AI frameworks [60, 84, 86, 99, 100, 102], which are not transferable across cultural contexts [8, 118]. The non-portability of Western frameworks can lead to flawed data and model assumptions, evaluation methods that overlook culturally-specific axes of discrimination, and cultural incongruencies [86, 101, 102]. When operationalized in model testing and evaluation, exclusive use of Western-oriented frameworks risks development of applications that dispossess the identity of non-Western communities [76], by centralizing the epistemologies used and power to build algorithmic systems in the hands of a global minority [50]. Compared to other AI harms, such as *representational* or *allocative harms*, much less attention has been devoted in computing literature to understanding cultural harms, leaving these "under articulated" in the field [105, p. 18]. Current approaches to understanding cultural harms focuses on how they can foreclose ways of understanding the social world [95], leading to systemic erasure [38], proliferating false ideas about cultural groups [100], and exporting Western ideas to the Global South [76]. However, the nuanced ways these take shape for different non-Western communities are not well-understood.

More globally inclusive and community-centered approaches to AI fairness and cultural harms require recontextualizing data and model evaluation — with an explicit incorporation of contextual axes of discrimination [102]. Particularly there are calls to meaningfully center different global communities and institutions in knowledge production processes [5, 104] and incorporate participatory practices that allow production of ML frameworks by impacted communities. [102]. Combined with community-centered research, ML practices that center reciprocity, reflexivity, and empowerment can help reshift power dynamics between technologists and marginalized communities [16, 61].

2.3 Regimes of South Asian Representation

Representation, such as through visual media, is the process of creating and communicating meaning about the social world [48]. There are no "true" representations; rather representations are "historically determined [social] construction(s)…mediated by social, ideological, and cultural processes" [36, p. 115]. The power to represent communities in ways that shape how they are understood can be understood as a *regime of representation* [48], a dominant system of media discourse, symbols and images that create particular narratives about already marginalized groups. Hall shows how these representations are not just one off instances but are part of a broader "regime of power", that is upheld across media systems, controlling and shaping how others see specific groups. In the Asian context, a dominant regime of representation is *Orientalism*, which refers to a broader system of thought, way of writing, and studying the "Orient," or Eastern world, that emerged in the 19th century [98]. As framed by Western geopolitical forces, the "Orient" became a singular stand-in for the numerous cultural and national boundaries of the Asia continent [24, 57]. Through Orientalism's *outsider's gaze*, the West imposed demeaning cultural stereotypes onto Asia: backwards, silently different, passive, and sexualized [24]. Creating harmful representations of Asia was important for "dividing up the difference between (Europe, the West, 'us') and the strange (the Orient, the East, 'them')" [98, p. 43]. Thus, Orientalism was not simply a way of thinking about South Asia, but a means to conceptualize the geography of the colonial world that made Asia susceptible to certain kinds of control and geopolitical management [24].

Orientalist tropes continue in contemporary media about South Asia. These including essentialist representations of the sub-continent as diseased and mentally ill [26, 40], impoverished [25, 37], and economically dysfunctional [13]. Reductive depictions of Asian women as sexually available and exotic [71, 72], or lacking agency through Western understandings of veiling [71, 109] are also common. While the range of Orientalist representations vary, they are united in distorting the meaning of a cultural practice or symbol through reduction and simplification [78].

Regimes of representation are important sites of analysis because they shape hegemonic ways of seeing and knowing about a culture or community, both externally and internally [69, 70]. Moreover, the reductive stereotypes, miscategorizations, and forms of erasure [70] can "block the capacity of marginalized groups … to imagine, describe, and invent [themselves] in ways that are liberatory" [53, p. 2]. To date, little is known about what regimes of representation T2I models contain and perpetuate, particularly as defined by South Asian communities. Recognizing these regimes is necessary to disrupt their harmful impacts.

3 METHODOLOGY

In alignment with broader calls for developing non-Western and community-centered responsible AI practices [76, 100], we engaged

participants from three South Asian countries through focus groups and a survey to: (1) collaboratively develop culturally-specific text prompts, (2) collectively reflect upon images generated by T2I models in response to culturally-specific text prompts, and (3) understand participant experiences of generated imagery. In this section, we provide context and details about our methodology. For additional details, please see Appendix **??**.

3.1 Site of Study: South Asia

We focus this study on South Asia, reflecting the cultural expertise of the lead author. As our goal is to localize understandings of T2I cultural failure modes, we recruited from three different South Asian nation-states: Bangladesh, India, and Pakistan. We recognize South Asia is a rich and complex region with many diverse cultures that could be subdivided along multiple other axes (e.g., gender, religion) [20]. The chosen nation-states share cultural histories with enough common overlap to facilitate a cohesive yet nuanced analysis and allow for more analytical comparisons.

3.2 Study Participants

We recruited a purposive sample [82] of participants with cultural knowledge of any one of the three nation-states in the form of lived experience, professional affiliation, and/or academic study about Pakistan, India, or Bangladesh. Purposive sampling was used to ensure diversity across the source of cultural knowledge and nation-state. We asked prospective participants to self-identify with one of the nation-states and describe their experience with the selected nation-state; we did not exclude participants based on citizenship or current residency. This allowed us to capture prospective participants with deep cultural knowledge in the South Asian diaspora who are living outside their home countries. Other inclusion criteria required participants have English-language proficiency and be at least 18 years old.

We recruited through targeted emails to (1) academic listservs focused on South Asian Studies programs in North America, Europe, and Asia registered with the Association for Asian Studies, as well as computing research listservs to recruit potential participants with diverse domain expertise directly relevant to our research aims; (2) cultural institutions in Pakistan, India, and Bangladesh; and (3) through the research team's professional networks in South Asia. We received 219 responses. We excluded those who had only visited a country for tourism, currently work for a major technology firm, and "spam" replies with misidentified provinces or languages. We invited 52 eligible participants and; 36 people ultimately participated (Pakistan (n = 15); India (n = 13); Bangladesh (n = 8). While we did not systematically recruit for intra-national cultural knowledge (e.g., linguistic and regional diversity), our sample covered ten linguistic groups and fourteen sub-national regional groups within South Asia. We had participant diversity across occupational expertise, including 17 academic researchers, 9 "cultural workers" employed in cultural industries, such as museum curation and the arts, and 10 participants with lived experience of the cultural contexts we were studying, but not necessarily professional experience in cultural industries. All participants received a localized equivalent of $300 USD in thanks for their participation.

3.3 Method of Engagement: Focus Groups

We crafted a study design that facilitates collective engagement and conversation: focus groups. For studying culture, focus groups offer an effective means of accessing culturally-specific knowledge [31, 91] and can "facilitate culturally sensitive research" [54, p. 777], as the setting helps foster cohesiveness among participants [67], where interactions among participants generate important information [115] on cultural representation and cultural harms [81]. We also created opportunities for anonymous feedback through digital whiteboards.

Each participant attended two 90-minute focus groups composed of between 7 and 9 participants from the same country to facilitate rapport [65] and allow participants to focus on the histories and cultures most relevant to them. The first focus group was structured around discussion questions and interactive activities to understand how participants defined "good" and "bad" cultural representations they had encountered in different media and how participants might assess "good" and "bad" representations in AI-generated imagery. To orient participants around the capabilities of T2I models we shared sample generated images of Western and South Asian cultural subject matter. The cross-cultural images served as a point of comparison in participant reflections. Following the first focus group, participants completed a survey submitting full-sentence text prompts and suggesting up to five examples of cultural events, landmarks, art styles and/or artists, historic events, figures, and characters they felt would enable the assessment of T2I models.

During the second focus group, participants reviewed generated images from 4-5 prompts, seeing four images per prompt. While this number of images is not sufficient to draw inferences about statistical distributions in T2I model outputs, the smaller sample enabled participants to conduct deeper reflections on generated images in alignment with study goals. During these reflections, we requested participants specifically identify what they thought generated images "got right" and what models did "poorly." Following individual reflection, we facilitated discussion on the possibilities and risks of T2I models. We deliberately kept discussion questions open–ended to give participants agency to focus on what they found most important.

3.4 Developing Prompts and Generating Images

Between the first and second focus group, the research team synthesized participant's prompt suggestions, tested various prompts, and generated images using four state-of-the-art T2I models [88, 92, 97, 122]. We constructed prompts based on participant suggestions with the aim of increasing quality and coverage of cultural references in the study. As the study's focus is on cultural limitations, we utilized prompts that minimized the likelihood of non-cultural failures to ensure we made best use of participant's time and expertise. For example, a general model failure for the prompt, "A day in Lahore," might result in images of daylight, rather than the city; however, rewording the prompt as "People spending their day in Lahore" led to images reflecting the model's learned associations with daily life in Lahore.

The final corpus included 120 prompts. We randomly assigned each prompt to two of four state-of-the-art T2I models for image generation. We then selected the first two images from each model

for participant feedback in the second focus group. Each participant was assigned to review the selected images from 4-5 text prompts. This scoping allowed for model coverage in the study, while keeping the number of image outputs manageable for participants to comment on during the focus group.

3.5 Data Analysis

All focus groups were video-recorded, transcribed, and thematically analyzed [21, 23]. We also compiled the video conference chat and participants' written feedback into the data corpus. All four authors participated in data analysis, which involved iterative and collaborative development and discussion of codes and themes, drawing upon the reflective thematic analysis approach described in [21, 23]. After reviewing and developing initial codes for all the focus group data, the research team shared, arbitrated, and iterated on codes, developing preliminary themes during a series of data sessions [22], aligning on themes that cut across the focus groups while attending to poignant differences between the cultural centers of focus. Interpreting raw data with relevant cultural scholarship textured participant contributions [113] with broader scale explanations and implications of cultural phenomena that participants could only briefly reference at the focus groups given time constraints. We introduce participant quotes with alphanumeric identifiers, providing their country to better contextualize participant comments. Some quotes were provided anonymously during interactive exercises, and they will only have the identifier "A" after the country-group.

3.6 Limitations

While our work offers important insights on cultural failure modes of T2I models, our methodology has limitations. A purposive sampling strategy and focus on three nation-states within a diverse region does not necessarily lead to insights that generalize across South Asia, especially as we did not systematically recruit for participant diversity, such as caste, ethno-linguistic identity, and class. Having English language proficiency as inclusion criteria, and participant and researcher academic affiliations also suggest participants may largely reflect upper-class and urban-dwelling subsets of South Asia. Finally, for time and safety reasons, we did not generate images with participants in real-time; and thus, researchers remained arbiters of shown images. In this way, we characterize this study as community-centered rather than participatory. Our process of prompt synthesis and image generation did not systematically disentangle the effects of different words or phrases on generated images, limiting our ability to draw inferences regarding why certain representations were emerging. We leave this important research direction to future work.

4 FINDINGS

4.1 Regimes of Representation in T2I Models

In this section, we outline three model failures foregrounded by South Asian participants and how they characterized these failures as reinscribing dominant *regimes of representation* [48]; in this case, the hegemonic ways of seeing, knowing and representing South Asia they experience. Broadly, participants were interested in both the accuracy of cultural subject matter recognition and nuances of cultural representations in T2I generated imagery. As P35, from Pakistan, summarized *"[if I put in a particular figure, historical event, or allegory], does [the model] get what I'm trying to say, first of all? Is there a kind of understanding or legibility? But then within that, what kind of visual representation do you get? Do you get a kind of Orientalist, portraiture rendering? Do you get an image that looks closer to maybe South Asian renderings?"*

Drawing from this call for cultural recognition, we present three failure modes that encapsulate participant concerns about model accuracy and representations: (1) failing to recognize cultural subjects: generated imagery fails to depict a culture's subject matter; (2) amplifying cultural defaults: culture's subject matter in generated images defaults to particular hegemonic cultures; and (3) perpetuating cultural tropes: generated images contain stereotypes and tropes associated with particular cultures.

4.1.1 Failing to Recognize Cultural Subjects. Participants shared their desire to test T2I models' ability to generate cultural artifacts, history, and practices from South Asian cultures. Importantly, participants were not looking for absolute accuracy in each image, and emphasized its impossibility for topics with multiple realities and possible renderings (e.g., a South Asian family). Rather, they adjudicated accuracy based on whether the cultural subject matter had a canonical rendering (e.g., historical figures like Indira Gandhi and architectural landmarks like Badshahi Mosque), or essential canonical elements (e.g., the correct sporting equipment for cricket scenes, the proper landscape for a region, or the art style of Sadequain). Participants emphasized their concerns about accuracy extended to equitable cross-cultural performance, as P17 from India explained: "*if [the machine] can recognize the style of Picasso, then, you know, is it equally possible for a machine to recognize the style of Warli paintings [a style of Indian folk art]?*" Through their reviews of generated imagery, participants identified different dimensions of "failure to recognize cultural subjects," from total failure to partial legibility lacking cultural specificity.

Across all countries, South Asian participants identified examples where models completely failed to depict important cultural subject matter specified in text prompts. For instance, models totally failed to render the styles of famous artists from India (e.g., Tagore), Pakistan (e.g., Gulgee, Sadequain), and Bangladesh (e.g., Zainul Abdein). Participants described how such total failures were particularly frustrating as generated images shown during the first focus group reflected the painting styles of Monet, Picasso, and Rembrandt in easily recognizable ways. These cross-cultural failures were obvious, as P18 from India reflected: *"AI seems to be able to pick up and adapt [images] to the style of Monet [...] much better [than with] Indian artists or Indian folk art."* Participants across all three regions also commented how well-known Western cultural figures, such as Sherlock Holmes, generated coherently with the visual markers typically associated with these figures.

Participants named a second way T2I models fail to recognize cultural subjects, in which models render vaguely "Eastern" visual associations in generated imagery. For example, a text prompt for the famous love story, Heer Ranjha, resulted in depictions that according to P30, a Pakistani participant, "*[do not] really have anything to do with Heer Ranjha*" (see Figure 1). The famous folklore story is about two star-crossed lovers from rural Punjab; however none of the generated images contained Heer, the woman, and included

only a man wearing attire completely disconnected from the Punjab region or class that Heer Ranjha were from. Explaining further, P30 described the man as a *"[stereotypical] monarch from Northern India,"* while Ranjha was a character from an agricultural family. Participants called out other failed images with vaguely "Eastern" aesthetics, including those generated from text prompts for specific South Asian cities that produced generic cityscapes or cities with inaccurate cultural markers. Reviewing generated images for the text prompt, "Children eating fried street food in Varanasi," P20 from India commented "*there is nothing recognizably Varanasi about those images, (...) this could be just be any generic, small town.*" For text prompts referencing Mughal-era (a South Asian empire with a distinctive style) cityscapes and buildings, models generated architecture that participants described as clearly Ottoman-looking, Gulf or Middle Eastern, or even East Asian-like, which created cultural incoherence in the generated images and indicated these cultures could be merged into one indistinguishable category.

Speaking to intra-cultural regimes of representation, participants were differentially satisfied with generated images of the three South Asian contexts. Generated images for North Indian artifacts, such as culturally important buildings like the Qutub Minar and Red Fort, were identified as more accurate than their Pakistani and Bangladeshi counterparts, such as the Baitul Mukarram National Masjid in Bangladesh. However, even within India, participants emphasized T2I models generated imagery more effectively for majority cultural artifacts compared to regional South Asian celebrations, such as Rajwadi Holi, which did not render at all.

Whether generated images completely failed or offered non-specific renderings, this failure mode speaks to the uneven performance of T2I models in recognizing different cultural subject matter producing unequal *quality-of-service* for different communities based on identity [15, 75] and cultural harms already present in mainstream media, such as whitewashing and Asian erasure [80].

4.1.2 Amplifying Hegemonic Cultural Defaults. Cultural defaults refer to which cultural centers are naturalized as the dominant frame of reference. As a T2I failure mode, cultural defaults encapsulate participant concerns about which cultural lens dominates representations in generated imagery. The overrepresentation of Western or white cultural subject matter in media and algorithmic technologies is now expected by scholarship (e.g., [2, 3, 42, 80]). Participants, too, mentioned white, Western defaults in media and identified examples where T2I models appeared to default to Eurocentric cultural artifacts, even if no cultural context was specified in the prompt. For example, neutral prompts for "A photo of a house of worship" rendered Christian, American-looking churches (see Figure 2) and "Toddler in marketplace" resulted in multiple images of white-skinned toddlers in stereotypically Western grocery stores. More worryingly, this centering of white, Western bodies continued even when South Asian cultural contexts were specified in text prompts (e.g., "Children eating street food in Varanasi," "People eating street food in Lahore," and "People celebrating Holi"), as illustrated in Figure **??** in Appendix **??**. However, participants went beyond the expected patterns of centering whiteness and Western culture to name a more complex *hierarchy of cultural defaults* they saw co-produced through T2I imagery: regional power centers and intra-national axes of discrimination.

Participants from Bangladesh and Pakistan, in particular, emphasized the ways their cultural identities are erased and miscategorized in dominant media centering India as the South Asian cultural default. To test if this cultural erasure extends to T2I models, some participants incorporated culturally-specific language in their prompts, such as the term "Deshi," which specifically refers to Bangladeshi people as opposed to the term used in Pakistan and India: Desi. For participants, generated images for "Deshi" felt more akin to Rajasthani (Indian) depictions. One anonymous comment summarized the significance of regional cultural erasure that "*India should not stand in for all of South Asia*", and P30, from Pakistan, explained how wide-sweeping this cultural default is, as South Asia is "*an area that is about as big or half as big as Western Europe. That is a very large area, and there are tons of cultures … generalized into Northern India.*" Bangladeshi and Pakistani participants commented how India as a regional cultural default emerged through T2I models, identifying a pattern among images generated from prompts referencing simply, "South Asia," which defaulted to what they viewed as Indian portrayals. For the prompt "South Asian family," Pakistani and Bangladeshi participants referred to generated attire as "Indian-looking;" similarly for prompts referencing women in "saris," Bangladeshi participants emphasized models produced saris with patterns and styling that appears Indian.

Participants also identified where T2I models generated Indian objects from prompts explicitly mentioning Bangladeshi and Pakistani cultural artifacts and subjects. For instance, prompts for Bangladeshi cultural topics rendered imagery containing Hindi Sanskrit instead of Bengali script. A prompt for "Bangladeshi Language Day" resulted in images with Hindi text; and images generated for the "Bangladesh Liberation War," a seminal moment in Bangladesh's history that formed the nation, depicted men wearing turbans, which P8 felt "*represent[ed] more an Indian army man than an actual Bangladeshi army.*" Reflecting on Bangladeshi cultural erasures, P2 emphasized: "*they didn't really get the exact nuances of our region or our people*" and P3 commented "*AI still has a lot to learn about South Asia, apart from India.*"

Beyond the "India as South Asia" cultural default, participants spoke to intra-national regimes of representation that erased the "*diversity of class, religious, gender, ethnic minority narratives*" (anonymous, Pakistan) within their countries; a pattern they felt T2I models replicated. When prompts explicitly mentioned India (e.g., "Indian food" and "Indian women"), generated images defaulted to what participants identified as upper-caste, North Indian cultural subject matter (see Figure **??** in Appendix **??**). When discussing images generated in response to the prompt "Indian cultural dancers," P15 identified the predominance of upper-caste dance forms, like Bharatnatyam, but not folk dancers of more marginalized castes, characterizing the representations as having a "*very homogenized lens (upper caste).*" She also pointed out that while dance forms practiced by men (bhangra) were represented in the images, women's dance traditions (e.g., giddha) were not, emphasizing generated imagery perpetuated a "*very male perspective through which we look into dance form.*" Religious diversity was also missing in most outputs for prompts referencing "Indian houses of worship." P12, from India, pointed out how they felt this was a "*Hinduization of Indian religious iconography*" in T2I imagery that mapped onto a braoder imaginary of India as unequivocally "Hindu" (see: [57]),

Figure 1: Example of a DALL-E generated image for the prompt "Heer Ranjha" (left) juxtaposed with a canonical representations of Heer Ranjha [56] (right) showing the generated image resembles a monarch or warrior instead of a couple from rural Punjab.

Figure 2: Generated images, from Imagen and Parti, for prompt "A photo of a house of worship" showing Western-looking churches.

Figure 3: Generated images, from Stable Diffusion and DALL-E, for prompt "People spending their day in Peshawar" showing dusty streets and markers of poverty and none of Peshawar's rich cultural heritage.

even though India has a significant minority of Muslims, Christians and Buddhists.

The "amplifying hegemonic cultural default" failure mode reinscribes existing regimes of representation, such as through miscategorization and homogenization. This includes Orientalist representations that continue to homogenize South Asian culture that render invisible inter- and intra-cultural differences across Asia and internal power centers speaking on behalf of marginalized communities.

4.1.3 Perpetuating Cultural Tropes. Cultural tropes reflect the stereotypes associated with particular cultures. Whereas the prior failure modes reflect systematic absences and miscategorizations, cultural tropes concern the harmful, essentialist representations that appear when cultural subject matter is visualized. These representations are "*caricatures of the world*" (P33, Pakistan) that perpetuate "*extremely narrow depictions of extremely diverse phenomena/lives that then come to stand for the whole*" (anonymous, Pakistan). Here, we summarize four dominant cultural tropes identified by participants, emphasizing connections between T2I imagery and existing South Asia regimes of representation.

South Asia as impoverished and under-developed. Participants across all three nation-states described how tropes of dusty cities and "*everyone living in slums*" (anonymous, India) pervade media portrayals of South Asia, reducing the region to "*one economic strata*" (anonymous, India). While income inequality exists, as with all parts of the world, the rich diversity of South Asian life is absent in media tropes depicting South Asia as unequivocally impoverished [25, 37] and economically dysfunctional [12]. Participants identified how this trope appeared in images generated from prompts for daily life in South Asian locales, often depicting "*shabby and old households*" (P4, Bangladesh). P21, from India, described images for, "A photo of daily life in Mumbai," reduced the city to "*congested spaces and poverty.*" For the prompt, "People spending their day in Peshawar" (see Figure 3), P30 from Pakistan emphasized how inclusion of architectural details would have disrupted the trope of underdevelopment: "*Peshawar has markets, (...) old frescos, (...) old buildings, the old woodwork from the pre-independence era. It has various cultural stalls. So... I would have wanted ... something that ... presents our culture (...) What I received was a dusty street with a few rickshaws.*" P9, from India, commented how generated images framed indigenous South Asian tribes as dirty "*even though Adivasi villages and homes are really clean and beautiful, even if there is poverty*" (see Figure **??** in Appendix **??**). Participants also noted how renderings represented South Asia as frozen in time, indicating it was less modern or advanced. Reviewing prompts for scenes and marketplaces in various Pakistani cities, P22 noted generated images erased "*modern parts of urban life,*" by showing only "*old school open markets,*" rather than the contemporary "*upscale marketplaces.*" In sum, participants felt they were "*seeing pictures [from] 50 years back*" (P31, Pakistan).

South Asia as exotic. Participants noted the harmful cultural trope of exoticization in media, which from a Western gaze, depicts South Asia as a strange and bizarre land [6]. Exocitization is a regime of representation meant to set South Asia as a land apart, and different from the West, something "out there." P20, from India, described how South Asia is imagined as having "*chaotic traffic*" and "*the cows in the streets,*" creating a representation of South Asia as disorderly and overpopulated. P12 mentioned the trope of India as a "*land of snake charmers,*" where South Asian men are depicted as excessively brown-skinned and women clothed in traditional attire. Others noted the association of South Asia with particular color palettes sets the region apart from the rest of the world — either sepia tones or over-the-top bright colors — constituting another form of exoticization in the media [68]. P11, from India, reflected how exoticization was common on postcard images depicting tribal women wearing "*extensive silver jewelry*" and positioned South Asian women as "*exotic and wondrous and magic*" subjects.

Participants identified this theme of exoticization in T2I imagery. In response to the prompt, "Painting of Queer South Asia where the painting has symbols of South Asia and queer culture," multiple participants noted generated images continued the trope of South Asia being reduced to a certain color palette. P36, from Pakistan, called these colors "gaudy," and P20, from India, pointed out that for Western representations, the models had "*a greater variety, a greater*

diversity of color palettes and styles." P36, from Pakistan, specifically called out the similarity of T2I outputs to historic photography practices, particularly colonial imagery: "*the way the darkness of these bodies is represented is uncannily like especially that the first hundred or so years in photography, when lighting and color and picture development processes were very unsurprisingly set towards representing white bodies more than dark bodies. So it just brings up that particular history in showing these ill-defined generic dark bodies, even if it's a little bit I guess more artistic.*" P11, from India, emphasized the invocation of such tropes is merely a way to sell more media, a capitalist and colonial logic that can continue in T2I, noting: "*And I think that should stop.*"

Dalit communities as disempowered. One pernicious trope identified in Indian media and T2I representations concerned Dalit communities as disempowered: a caste group in India, mapping on to the lowest rung of the caste hierarchy who have experienced centuries of social and economic marginalization and exclusion [90]. P15, a Dalit academic from India, discussed how Dalit communities are often presented in the media through both a classist and casteist lens, associated with "undesireable" occupations: "*[near] a sewer or toilets... with dirt around [them].*" Reviewing a prompt for "Daily life of a Dalit person," she pointed out T2I imagery similarly associated Dalit life with connotations of dirtiness, hardship, poverty, and lacking artisanal and resistive culture. None of the generated images for "daily life" incorporated Dalit celebrations or cultural productions, such as Dalit dance forms. Even when models were prompted for "A Dalit family celebrating Diwali in their house," P15 detailed how the model resorted to upper-caste Hindu celebrations of Diwali that did not show the specific characteristics of Dalit Diwali celebrations. She further emphasized that representations of Dalit daily life should "*also [be] about their songs, about their cultures, about how they make a difference through their everyday acts.*" Characterizing T2I imagery as "*essentialist*" and a "*cliched representation,*" P15 specified this regime of representation missed the "*dynamic aspect of Dalit identity*" that in reality, disrupts the trope of "*abject poverty…as the only marker of Dalit life.*"

Muslim lives as one-dimensional. Pakistani participants expressed frustration with Western media narratives that reduce Islam and Muslim cultures to religious iconography, which in the post-9/11 era portray Pakistan as a "terrorist" nation [73] and Islam as fundamentalist [63]. On media narratives, P23 from Pakistan reflected: "*When we talk about Muslim life it always goes with a mosque. [As if Muslims] only worship all day.*" Similarly, P33 from Pakistan described a fixation on "*the call to prayer at the beginning of all TV shows and movies.*" Participants discussed nuanced ways these tropes appeared in T2I imagery. For instance, through repeated depictions of people wearing traditionally religious attire in scenes of Pakistan. In response to the prompt "Political protest in Pakistan," P26 noted: "*All men are wearing shalwar kameez and most are wearing prayer caps. Literally no person is wearing Western attire, which is quite common for men in Pakistan.*" P23 and P22 both talked about the constant presence of veils and headscarves in T2I depictions of Muslim women, which P26 noted mapped onto tropes of women as only "*conservative;*" a trope that communicates Muslim women need rescuing [58] and lack agency [1, 110] (see Figure **??** in Appendix **??**). Participants clarified when prompts specify religious subject matter (e.g., Eid), veils and headscarves are not inherently problematic; but their presence in all generated images speaks to how Muslim lives are condensed to one-dimensional stories. T2I models risk reproducing reductive and one-dimensional representations for complex cultural subjects, such as "Islam," further reducing diverse Islamic cultures to one form of religious practice. For example, participants described depictions of an "Islamic city" as being "*very stereotypical*" (P24, Pakistan) due to the images' fixation on mosques. P23 noted the reduction of Muslim cities to mosques made it seem like people in these cities "*don't have a life*" outside religion.

4.2 Negotiating Outsider Gazes

The failure modes our participants identified map onto existing power structures and logics of power for representing South Asia, pointing to how T2I models can perpetuate multiple "outsider gazes." In this section, we connect model failures to the social and political dynamics participants experience in their lives and present participant discussions on the possibilities of inclusive and representative AI systems.

4.2.1 Cultural impacts. The T2I limitations participants identified have a long history in media representations of South Asia, where "the touristic, Western gaze" is pandered to [27, p. 7]. Participants expressed concern that T2I models might be heavily biased towards outsider perspectives on their cultures. P9, from India, described how generated images felt like "*tourist's photos*" reflecting "*flatter versions of South Asia*", and amplified what P17, from India, called the "*empirical abundance of certain kinds of images [about India]*" that map onto global majoritarian views. P30 described they felt T2I training data and the resulting imagery led to a "*Western vision of the East.*" However, even within South Asia, those with greater social power can produce representations of marginalized communities that are just as "othering" as those produced by the West. As P13, from India, commented, "*it's not just South Asian culture here against … Western culture… There's so many layers here of hegemonic cultures within South Asia. One small layer of this [culture] gets to represent the entirety of South Asian culture.*" Participants complicated the idea of "South Asia" by discussing the internal power centers that reproduce colonial representations of the region, reifying exclusionary and problematic cultural representations. Reflecting on the layers of hegemony represented in generated images, P19, from India, commented "*[AI] keeps privileging so much that has been privileged. AI keeps amplifying the privileged voice.*"

Participants described how they negotiate and work to correct reductive media stereotypes in their lives, and were concerned T2I models would further "normalize" and give authenticity to these stereotypes. P23 and P26, from Pakistan, explained how media depictions that reduce Muslim culture to religious rituals create tension and awkward social moments when they traveled outside Pakistan. P12, from India, described how when she travels abroad, she is frequently asked if "*India is full of slums like in [the film] Slumdog Millionaire.*" Participants reflected on the frustration and grief related to identity loss when their cultures are conflated with others. P8 elaborated that, as a Bangladeshi, such points of cultural confusion are regular occurrences: "*Growing up, I was always categorized as … [Indian] I'm like no, I'm not Indian… No, I'm*

Bangladeshi... we have our own foods, we have our own holidays, we have our own historical events." Similarly, P4, from Bangladesh, voiced concern that people in the Bangladeshi diaspora growing up outside their country would lose their cultural identity because they are less attuned to these differences. P22, from Pakistan, described the distress of seeing outsider representations of their culture in T2I outputs: "*AI represents the majoritarian view and if you're someone who doesn't fit in with that, then it's particularly disturbing [for you].*" Generative AI that reify cultural power hierarchies risks limiting how people understand their culture on their own terms. If empowering cultural representations are not reflected in emerging media, including AI technologies, "*we stop imagining ourselves to be what we are*" (P28, Pakistan). When algorithms reproduce and amplify an outsider's narrative about a culture, they impact both people's sense of identity, belonging and how they are perceived by others as a form of *algorithmic symbolic annihilation* [64].

4.2.2 Aspirations for Generative AI. While participants agreed on the importance of T2I models being inclusive of global cultures, multiple participants emphasized the challenges inherent in defining and operationalizing global inclusively. As P14 from India explained, "*there's no singular identity,*" but rather "*multiple languages, multiple cultures [...] and the complexities that come with that.*" Participants also commented on the subjectivity inherent in the interpretation of visual imagery, echoing AI scholarship that has written of the socially and culturally subjective nature of image-text relationships [55]. As one participant put it, most text prompts will have "*such a wide range of portrayals*" that there will always be the question of "*which lens are you using?*" (P26, Pakistan) This echoes literature on representation that argues representations are always "'positional truths' which are linked to history, power, and dominance within a global context mediated by economic, political, ideological, and cultural processes" [36].

While pointing out the limits of T2I models, participants shared nuanced perspectives on the potential they saw generative AI could have in challenging outsider gazes and power inequities in existing archives and media. They noted new sources of media could bring out possibilities of multiplicity and diversity of representations unconstrained by existing hierarchies of caste, class, or museum patronage. They pointed to, for instance, the diverse representations generated by South Asian communities on TikTok, commenting there was "*already an overwhelming incredible diversity of visual vocabularies and modes [available online],*" that we could learn from instead of having our "*representation strained by logics of power and capital*" (P36, Pakistan). Participants discussed whether generative AI could grant people space to tell their own stories and represent themselves, seeing an opportunity for generative AI to "*call attention to certain kinds of folk art practices, which otherwise nobody would have noticed*" (P17, India).

However, participants questioned the bounds of what should be captured in T2I models, debating the values and risks of inclusion; and raising concerns about artist attribution, commodification, and the consequences of separating certain art forms from their traditional roots. For example, when reflecting on the models' failure to produce a kalamkari-style print, P14 from India argued that the easier it becomes to "*find traditional artworks that [have been] mass produced somewhere [...] the more it [becomes a] mechanism to run roughshod over people's practices that don't already have a voice and just further silence them or push them into obscurity or further.*" Ultimately, participant aspirations for generative AI focused heavily on restoring agency and community-control over the terms of representation, exemplified by one participant's challenge: "*why can't we imagine a more authentic world that our communities can build ourselves*" (P32, Pakistan). P19 from India similarly argued for shifting the power imbalance by "*including people in this process... [as] representation has to come from places, which do not or never have had the resources to tell the stories.*" Other participants were less hopeful about the possibility of inclusion in AI, instead questioning whether "*it's better to just opt out*" because: "*there's no way for this to be equitable. The bulk of material, the weight of how much media has already been produced, the sheer volume of it is so huge that it's never going to be representative*" (P22, Pakistan).

5 DISCUSSION

In this section, we offer provocations to computing researchers for building a research agenda for globally responsible generative AI, while recognizing questions of inclusion and representation are not straightforward and deserve thoughtful attention. We argue for "culture" as an analytic for research on generative AI and reflect on the importance and complexity of community-centered research.

5.1 Cultural Power and Emerging Generative AI

Generative AI technologies are increasingly producing cultural artefacts. Thus, as they are launched globally for global populations, they are inevitably contributing to and situated within existing circuits of global cultural power. For instance, as with other media and technologies of cultural preservation and representation, generative AI will also have to contend with whose cultural narratives get reproduced, and whose cultural knowledge is erased through these technologies. [48]. Similarly, we have to consider who has the power to tell their own stories through these technologies, and which communities are not represented on their own terms [30]. To contend with and recognize this power within T2I models specifically — and generative AI broadly — we must undertake empirical studies of how algorithmic systems amplify, shape or are shaped by broader cultural relationships they are being launched within. This approach suggests investigating the cultural lineages of the digitized archives of cultural production, such as museum collections, that constitute the datasets T2I models are trained on. It also includes considering the ways in which model outputs can stabilize and scale existing regimes of representation through patterns of over- and under-visibility for populations around the globe [79, 120]. We also argue for expanding empirical studies to include examinations of the cultural harms and impacts of these technologies, for instance how they may contribute to cultural hegemony, cultural erasure, or cultural stereotyping. This approach de-centers the model or technology itself as the sole loci of impact, and focuses on the broader social and cultural milieu within which AI is produced, deployed and used to understand its performance and impact. In this, we build on works of scholars who study algorithms as culturally-produced and situated objects [29, 41, 46, 49, 103].

Apart from empirical research, T2I models, and other related generative image technologies, must be historicized within scholarly

analyses of other cultural technologies, such as how photography functioned as a technology of cultural memory, propagation, and inclusion and exclusion [9, 94]. Insights about the politics of cultural technologies, particularly how they functioned in historically exclusionary ways to certain communities [11], shed light on how generative image technologies may have complicated relationships with communities historically marginalized from canonical, majoritarian representations. Through such cross-disciplinary engagement, FAccT can strengthen its analyses by linking individual testimony to broader cultural experiences and collective social structures because model outputs become pernicious precisely because they amplify existing power inequities and dominant logics of knowing a culture. Building culturally-inclusive technologies requires learning from both communities and scholarship about possible harms inflicted under the guise of technological cultural inclusion.

5.2 Considerations for Community-Centered Research

Community-centered model evaluations are essential for building contextually sensitive evaluative criteria and strengthening harms frameworks so they are rooted in rich understanding of lived experiences. Participants in our study foregrounded specific ways caste, gender, religion, and occupation intersect across South Asia to produce social inequalities, sharing the material markers they use to identify axes of discrimination, from attire to architecture to occupations. This cultural knowledge adds empirical depth and specificity to calls by Global South scholars to contextualize axes of discrimination in our evaluative frameworks [101]. Whereas researchers can acquire knowledge about the kinds of inequalities different communities' experience [93, 108], directly engaging local communities' standpoints [51] can strengthen the nuance of findings. Participants in our study underscored how if "outsiders" to a culture, whether data annotators or researchers, tried to evaluate these images they would not have enough cultural knowledge to recognize the nuanced ways in which cultural subject matter was mis-generated. This concern echos an emerging body of work that identifies lived experience as a valuable form of expertise within data annotation pipelines [35, 39, 44, 47]. Developing AI harms frameworks through direct engagement with communities is critical; as one participant pointed out, researchers may overlook problematic cross-cultural representations, even if attuned to harmful representations in their own context. Unlike, crowdsourced annotator studies, direct community engagement provides an iterative space where community members have more agency to articulate AI harms in their own voice and not be constrained by pre-determined metrics or categories.

Large-scale systematic analysis of the prevalence of failure modes, and causal factors underlying them, is out of scope for this study. However, our study offers insights into how robust methods of detection and mitigation can be rooted in rich community-centered conceptualizations of cultural harm and failure modes. For example, we gained nuanced understandings of cultural failure modes tied to experiences of marginalization, participants' alternative imaginations of AI, and specific understanding of local axes of discrimination that can improve model evaluation and provide deeper cultural knowledge for contextual metrics and testing prompts. However, our study also represents a limited form of community-engagement. In future research, we hope to build participatory structures facilitating more equitable power sharing, such as co-designing the research agenda with communities and co-analyzing the results and mitigation strategies.

However, alongside the growing calls for more community-centered and participatory research in ML [85, 106, 112], critical scholarship cautions participation is not a panacea to historic power imbalances [33, 107]; we can not just simply add new stakeholders to an inequitable system [34] and expect the system to transform [52]. In conversation with these concerns, our study offers opportunity to reflect on the very constitution of "community." Within our study it was important to acknowledge communities are not natural sites of perfect inclusion [59] and are themselves mired in multiple, overlapping center-periphery dynamics. Recognizing intra-community dynamics of power are important in any structures of participation, and in particular here, as critical development studies and postcolonial scholars have long questioned the crafting of the 'Third World' as a homogenous group [32, 74, 77], with no "difference, hierarchy, and oppression within the invoked group" [59, p. xxiv]. Recognition of intra-community power dynamics is particularly important for evaluating cultural harms as community knowledge is not "a fixed commodity that people intrinsically have" [66, p. 17], but is produced through socially-situated and political processes [36] that are in turn shaped by these dynamics of power.

Our study reiterated how cultural standpoints can shape image evaluation as we saw how participants' social location such as caste, class, and ethnicity, influenced their interpretations of images and the harmfulness of representations they contained. Participants belonging to an oppressed caste in India identified and described in depth the disempowering tropes of Dalit representation, while others did not. Participants from Pakistan and Bangladesh emphasized the different kinds of "Indian-ness" of South Asian representations, something that did not come up as forcefully in the India focus groups. Participants' various disciplinary training and professional experience brought additional expertise to their judgments; as in our study, artists from different traditions recognized unique nuances to the stylistic, architectural, and artistic failures. As computing researchers increase their focus on the influence of sociocultural factors on annotation work [39, 47], it is essential to recognize how the situated knowledge and perspectives of annotators beyond demographic characteristics can impact image evaluation outcomes.

At the same time, we recognize the futility of what Miranda Joseph critiques as pursuing the creation of "more finely grained measures of authentic identity, producing not a critique of community but a proliferation of communities" [59, p. xxiii]. No definition of community will be perfectly inclusive, because boundaries by definition are exclusionary. Yet the act of creating a more granular community grouping (e.g. South Asian > Indian > Dalit) may occlude internal inequities and differences and present false homogeny. Any structures of participation we construct for community engagement can not aim for some perfect representation of all perspectives, but need to be vigilant about these intersecting forms of privilege and marginalization that influence whose voices are centered and excluded through our definitions of community, which in turn influences what is evaluated and how it is evaluated.

6 CONCLUSION

While this work aims to inform the development of culturally-inclusive generative AI, we do not wish to reinscribe incomplete notions of authentic or true representation, or cultural inclusion, knowing representation can never be complete. Nor do we want to rehash simplistic binaries of North/South, East/West. In fact, our study "up-ends" commonly held research practices that homogenize the South in opposition to the Global North. Instead, we argue for the need to consider our ideals and processes of inclusion within AI development. Knowing "the appearance of diversity is one thing, the implementation of meaningful diversity is another" [43, p. 99], we have to recognize that cultural limitations of generative AI are deeply entangled with structural and power inequalities, and we have to allow for that recognition within our AI development systems. We thus see this work as one step amongst many towards creating spaces of agency for communities to tell their own stories within and through AI. As P14, from India emphasized, the aim is not just "*tokenistic representation*" for communities, but foundational respect.

ACKNOWLEDGMENTS

We thank Vinodkumar Prabhakaran, Michael Madaio, Gurleen Virk, Kathy Meier-Hellstern, Sarah Laszlo, and the anonymous reviewers for their valuable feedback on the paper. We also thank our study participants for sharing their time and expertise.

REFERENCES

[1] Fauzia Ahmad. 2003. *South Asian Women in the Diaspora.* University of Pennsylvania Press, Chapter Still 'In Progress?' – Methodological Dilemmas, Tensions and Contradictions in Theorizing South Asian Muslim Women.

[2] Ali Alkhatib. 2021. To live in their utopia: Why algorithmic systems create absurd outcomes. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–9.

[3] James A Allen. 2019. The color of algorithms: An analysis and proposed research agenda for deterring algorithmic redlining. *Fordham Urb. LJ* 46 (2019), 219.

[4] Sareeta Amrute, Ranjit Singh, and Rigoberto Lara Guzmán. 2022. A Primer on AI in/from the Majority World: An Empirical Site and a Standpoint. *Available at SSRN 4199467* (2022).

[5] Chinmayi Arun. 2019. AI and the global south: Designing for other worlds. (2019).

[6] Vivek Bald. 2015. American orientalism. *Dissent* 62, 2 (2015), 23–34.

[7] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? *arXiv preprint arXiv:2210.15230* (2022).

[8] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 167–176.

[9] Ali Behdad and Luke Gartlan. 2013. *Photography's Orientalism: New Essays on Colonial Representation.* Getty Publications.

[10] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. *Proceedings of FAccT* (2021).

[11] Ruha Benjamin. 2020. Race after technology: Abolitionist tools for the new jim code.

[12] Manu Bhagavan and Faisal Bari. 2001. (Mis)Representing Economy: Western Media Production and the Impoverishment of South Asia. *Comparative Studies of South Asia, Africa and the Middle East* 21, 1-2 (08 2001), 99–109. https://doi.org/10.1215/1089201X-21-1-2-99 arXiv:https://read.dukeupress.edu/cssaame/article-pdf/21/1-2/99/402816/csa_21_1-2-13bhagavan.pdf

[13] Manu Belur Bhagavan and Faisal Bari. 2001. (Mis)Representing Economy: Western Media Production and the Impoverishment of South Asia. *Comparative Studies of South Asia, Africa and the Middle East* 21, 1 (2001), 99–109.

[14] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759* (2022).

[15] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[16] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Abeba Birhane and William Isaac and Vinodkumar Prabhakaran and Mark Díaz and Madeleine Clare Elish and Iason Gabriel and Shakir Mohamed. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO 22).* Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages.

[17] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV).* 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158

[18] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[19] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The Forgotten Margins of AI Ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22).* Association for Computing Machinery, New York, NY, USA, 948–958. https://doi.org/10.1145/3531146.3533157

[20] Sugata Bose and Ayesha Jalal. 2017. *Modern South Asia: History, culture, political economy.* Taylor & Francis.

[21] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[22] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[23] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.

[24] Carol A. Breckenridge and Peter van der Veer. 1993. *Orientalism and the Postcolonial Predicament: Perspectives on South Asia.* University of Pennsylvania Press, Chapter Orientalism and the Postcolonial Predicament.

[25] Sarah Brouilette. 2011. *Re-Orientalism and South Asian Identity Politics: The Oriental Other Within.* Routledge, Chapter On the entrepreneurial ethos in Aravind Adiga's The White Tiger.

[26] Jennifer Burr. 2002. Cultural stereotypes of women from South Asian communities: mental health care professionals' explanations for patterns of suicide and depression. *Social science & medicine* 55, 5 (2002), 835–845.

[27] Shohini Chaudhuri. 2009. Snake charmers and child brides: Deepa Mehta's Water, 'exotic' representation, and the cross-cultural spectatorship of South Asian migrant cinema. *South Asian Popular Culture* 7, 1 (2009), 7–20. https://doi.org/10.1080/14746680802704956 arXiv:https://doi.org/10.1080/14746680802704956

[28] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers. *CoRR* abs/2202.04053 (2022). arXiv:2202.04053 https://arxiv.org/abs/2202.04053

[29] Angèle Christin. 2020. The ethnographer and the algorithm: beyond the black box. *Theory and Society* 49 (10 2020), 1–22. https://doi.org/10.1007/s11186-020-09411-3

[30] Patricia Hill Collins. 1999. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment.* Taylor & Francis Group.

[31] Erminia Colucci. 2008. On the use of focus groups in cross-cultural research. *Doing cross-cultural research: Ethical and methodological perspectives* (2008), 233–252.

[32] William Cooke and U. Kothari. 2001. *Participation: the new tyranny? (eds.).* Zed Books, United Kingdom.

[33] Ned Cooper, Tiffanie Horne, Gillian R Hayes, Courtney Heldreth, Michal Lahav, Jess Holbrook, and Lauren Wilcox. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 73, 18 pages. https://doi.org/10.1145/3491102.3517716

[34] Fernando Pedro Delgado, Stephen Yang, Michael A. Madaio, and Qian Yang. 2021. Stakeholder Participation in AI: Beyond "Add Diverse Stakeholders and Stir". *ArXiv* abs/2111.01122 (2021).

[35] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation. In *Proceedings of NeurIPS 2021 Workshop on Data-Centric AI.*

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

[37] Jigna Desai. 2011. *Re-Orientalism and South Asian Identity Politics: The Oriental Other Within.* Routledge, Chapter Pulp Frictions.

[38] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating Users' Strategies for Uncovering Harmful Algorithmic Behavior. In *Proceedings of the*

2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 626, 19 pages. https://doi.org/10.1145/3491102.3517441

[39] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2342–2351. https://doi.org/10.1145/3531146.3534647

[40] Christine Doran. 2016. Popular Orientalism: Somerset Maugham in Mainland Southeast Asia. *Humanities* 5, 1 (2016), 13.

[41] Paul Dourish. 2016. Algorithms and their others: Algorithmic culture in context. *Big Data & Society* 3 (2016).

[42] Moa Eriksson Krutrök and Mathilda Åkerlund. 2022. Through a white lens: Black victimhood, visibility, and whiteness in the Black Lives Matter movement on TikTok. *Information, Communication & Society* (2022), 1–19.

[43] Augie Fleras and Jean Lock Kunz. 2001. *Media and minorities: Representing diversity in a multicultural Canada.* Thompson Educational.

[44] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency*.

[45] Songwei Ge and Devi Parikh. 2021. Visual Conceptual Blending with Large-scale Language and Vision Models. *CoRR* abs/2106.14127 (2021). arXiv:2106.14127 https://arxiv.org/abs/2106.14127

[46] Tarleton Gillespie. 2016. *#Trendingistrending: When Algorithms Become Culture.* Routledge. https://www.microsoft.com/en-us/research/publication/trendingistrending-when-algorithms-become-culture-3/

[47] Nitesh Goyal, Ian Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proceedings of ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)* (2022).

[48] Stuart Hall. 1997. *Representation: cultural representations and signifying practices.* Vol. 1997. Sage London.

[49] Blake Hallinan and Ted Striphas. 2016. Recommended for you: The Netflix Prize and the production of algorithmic culture. *New Media & Society* 18, 1 (2016), 117–137. https://doi.org/10.1177/1461444814538646 arXiv:https://doi.org/10.1177/1461444814538646

[50] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512. https://doi.org/10.1145/3351095.3372826

[51] Sandra Harding. 2013. Rethinking standpoint epistemology: What is "strong objectivity"? In *Feminist epistemologies*. Routledge, 49–82.

[52] Anna Lauren Hoffmann. 2021. Terms of inclusion: Data, discourse, violence. *New Media & Society* 23, 12 (2021), 3539–3556.

[53] bell hooks. 1992. *Black looks: Race and representation.* South End Press.

[54] Diane L Hughes and Kimberly DuMont. 2002. Using focus groups to facilitate culturally anchored research. In *Ecological research to promote social change*. Springer, 257–289.

[55] Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. Underspecification in Scene Description-to-Depiction Tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online only, 1172–1184. https://aclanthology.org/2022.aacl-main.86

[56] Ibnazhar. 2016. *'Pakistan'- Islamabad Saidpur Village.* https://www.google.com/url?q=https://commons.m.wikimedia.org/wiki/File:%2527Pakistan%2527-_Islamabad_Saidpur_Village_-_@ibneazhar_Sep_2016_(141).jpg&sa=D&source=docs&ust=1682549051045601&usg=AOvVaw3CYcKTtwuHxJ0gFyD4xuv_

[57] Ronald Inden. 2001. *Imagining India.* Indiana University Press.

[58] Yasmin Jiwani. 2009. Helpless Maidens and Chivalrous Knights: Afghan Women in the Canadian Press. *University of Toronto Quarterly* 78, 2 (2009), 728–744. https://doi.org/10.3138/utq.78.2.728

[59] Miranda Joseph. 2002. *Against the Romance of Community.* University of Minnesota Press.

[60] Amba Kak. 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* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 307–312. https://doi.org/10.1145/3375627.3375859

[61] Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature. https://www.nature.com/articles/d41586-020-02003-2

[62] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% Right and Safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. https://doi.org/10.1145/3491102.3517533

[63] Karim H. Karim. 1997. The historical resilience of primary stereotypes: Core images of the Muslim Other. In *The language and politics of exclusion: Others in discourse*, Stephen H. Riggins (Ed.). Thousand Oaks, CA: Sage Publications, 153–182.

[64] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (2021).

[65] Jerome Kirk, Marc L Miller, and Marc Louis Miller. 1986. *Reliability and validity in qualitative research.* Sage.

[66] U. Kothari. 2001. Power, knowledge and social control in participatory development. In *Participation: the new tyranny?*, William Cooke and U. Kothari (Eds.). Zed Books, United Kingdom, 139–152.

[67] Richard A Krueger and Mary Anne Casey. 2000. *Focus Groups: A Practical Guide for Applied Research.* SAGE Publications, Thousand Oaks, CA.

[68] Jagjeet Lally. 2019. Colour as Commodity: Colonialism and the sensory worlds of South Asia. In *Third Text Forum online*. Taylor & Francis.

[69] Lisa Lau. 2009. *Re-Orientalism: The Perpetration and Development of Orientalism by Orientals.* Cambridge University Press.

[70] Lisa Lau and Ana Cristina Mendes. 2011. Introducing re-Orientalism: A new manifestation of Orientalism. In *Re-Orientalism and South Asian Identity Politics: The Oriental Other Within*. Routledge, 1–14.

[71] Reina Lewis. 2004. *Rethinking Orientalism: Women, travel and the Ottoman harem.* Vol. 4. Taylor & Francis.

[72] Sunaina Maira. 2008. Belly dancing: Arab-face, Orientalist feminism, and US empire. *American Quarterly* 60, 2 (2008), 317–345.

[73] Pavan Kumar Malreddy. 2015. *Orientalism, Terrorism, Indigenism: South Asian Readings in Postcolonialism.* Sage Publications.

[74] K McKinnon. 2006. An orthodoxy of 'the local': post-colonialism, participation and professionalism in northern Thailand. *The Geographical Journal* 172 (2006), 22–34. https://doi.org/10.1111/j.1475-4959.2006.00182.x

[75] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I Don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence* (2021), 169.

[76] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33 (2020), 659–684.

[77] Chandra Mohanty. 1988. Under Western Eyes: Feminist Scholarship and Colonial Discourses. *Feminist Review* 30, 1 (1988), 61–88. https://doi.org/10.1057/fr.1988.42 arXiv:https://doi.org/10.1057/fr.1988.42

[78] Brigitte L Nacos and Oscar Torres-Reyna. 2004. Framing Muslim-Americans before and after 9/11. In *Framing Terrorism*. Routledge, 141–166.

[79] Lisa Nakamura. 2007. *Digitizing race: Visual cultures of the internet.* University of Minnesota Press.

[80] David Oh. 2022. Whitewashing the Movies: Asian Erasure and White Subjectivity in US Film Culture.

[81] Anthony J Onwuegbuzie, Wendy B Dickinson, Nancy L Leech, and Annmarie G Zoran. 2009. A Qualitative Framework for Collecting and Analyzing Data in Focus Group Research. *International Journal of Qualitative Methods* 8, 3 (Sept. 2009), 1–21.

[82] Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research* 42 (2015), 533–544.

[83] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[84] Marie-Therese Png. 2022. At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1434–1445. https://doi.org/10.1145/3531146.3533200

[85] Vinodkumar Prabhakaran and Donald Martin. 2020. Participatory Machine Learning using Community Based System Dynamics. *Health and Human Rights Journal* (2020).

[86] Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural Incongruencies in Artificial Intelligence. *arXiv preprint arXiv:2211.13069* (2022).

[87] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research,*

Vol. 139), Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[88] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR* abs/2204.06125 (2022). https://doi.org/10.48550/arXiv.2204.06125 arXiv:2204.06125

[89] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

[90] Anupama Rao. 2009. *The caste question: Dalits and the politics of modern India*. Univ of California Press.

[91] Katrina L Rodriguez, Jana L Schwartz, Maria KE Lahman, and Monica R Geist. 2011. Culturally responsive focus groups: Reframing the research experience to focus on participants. *International Journal of Qualitative Methods* 10, 4 (2011), 400–417.

[92] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://github.com/CompVis/latent-diffusionhttps://arxiv.org/abs/2112.10752

[93] Hilary Rose. 1983. Hand, brain, and heart: A feminist epistemology for the natural sciences. *Signs: journal of Women in Culture and Society* 9, 1 (1983), 73–90.

[94] Daniel Rubinstein and Katrina Sluis. 2013. The digital image in photographic culture: Algorithmic photography and the crisis of representation. In *The photographic image in digital culture*. Routledge, 22–40.

[95] Jathan Sadowski and Evan Selinger. 2014. Creating a taxonomic tool for technocracy and applying it to Silicon Valley. *Technology in Society* 38 (2014), 161–168.

[96] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).

[97] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* (2022).

[98] Edward W. Said. 1978. *Orientalism*. Pantheon Books.

[99] Nithya Sambasivan. 2022. All Equation, No Human: The Myopia of AI Models. *Interactions* 29, 2 (feb 2022), 78–80. https://doi.org/10.1145/3516515

[100] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-Imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 315–328. https://doi.org/10.1145/3442188.3445896

[101] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 315–328.

[102] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, and Vinodkumar Prabhakaran. 2020. Non-portability of Algorithmic Fairness in India. *arXiv preprint arXiv:2012.03659* (2020).

[103] Nick Seaver. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4, 2 (2017), 2053951717738104. https://doi.org/10.1177/2053951717738104 arXiv:https://doi.org/10.1177/2053951717738104

[104] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[105] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2022. Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction. *arXiv preprint arXiv:2210.05791* (2022).

[106] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-Centered, Deliberation-Driven AI Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 440–451. https://doi.org/10.1145/3531146.3533110

[107] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is Not a Design Fix for Machine Learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 6 pages. https://doi.org/10.1145/3551624.3555285

[108] Dorothy E Smith. 1987. *The everyday world as problematic: A feminist sociology*. University of Toronto Press.

[109] Amira El-Azhary Sonbol. 2005. *Beyond the Exotic: Women's Histories in Islamic Societies*. Syracuse University Press, Chapter Introduction.

[110] Amira El Azhary Sonbol. 2005. *Beyond the Exotic: Women's Histories in Islamic Societies*. Syracuse University Press, Thousand Oaks, CA.

[111] Ramya Srinivasan and Kanji Uchino. 2021. Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 41–51.

[112] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 667–678. https://doi.org/10.1145/3531146.3533132

[113] Stefan Timmermans and Iddo Tavory. 2012. Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological theory* 30, 3 (2012), 167–186.

[114] Nenad Tomasev, Jonathan Leader Maynard, and Iason Gabriel. 2022. Manifestations of Xenophobia in AI Systems. *arXiv preprint arXiv:2212.07877* (2022).

[115] Sharon R Vaughn, Jeanne Shay Schumm, and Jane M Sinagub. 2012. *Focus group interviews in education and psychology*. SAGE Publications, Thousand Oaks, CA.

[116] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring Representational Harms in Image Captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 324–335. https://doi.org/10.1145/3531146.3533099

[117] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks Posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

[118] Lindsay Weinberg. 2022. Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research* 74 (2022), 75–109.

[119] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2022. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. *arXiv preprint arXiv:2212.11261* (2022).

[120] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 1–24.

[121] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* (2022).

[122] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. https://doi.org/10.48550/ARXIV.2206.10789