Beyond Categories of Caste: Examining Caste Bias and Morality in Text-to-Image AI Models

DIVYANSHU KUMAR SINGH, University of Colorado Boulder, USA
DIPTO DAS, University of Toronto, Canada
DEEPIKA RAMA SUBRAMANIAN, University of Colorado Boulder, USA
KOUSTUV SAHA, University of Illinois Urbana-Champaign, USA
STEPHEN VOIDA, University of Colorado Boulder, USA
BRYAN SEMAAN, University of Colorado Boulder, USA

Text-to-Image (T2I) models have shown promising utility across various domains. However, such models are also amplifying harmful societal biases in their outputs. In the context of South Asia, recent work has shown caste biases and stereotypes are being perpetuated through Generative AI (GenAI) systems. While this research offers extremely relevant insight into invisibilized narratives of caste discrimination through the GenAI system, they often treat caste as an identity category. Therefore, in this work we shift our ontology to focus on the relational aspect of caste. This enables us to develop a more nuanced understanding of the mechanics of caste discrimination by and through T2I models. Combining an algorithmic audit with critical discourse analysis, we draw on a conceptual frame challenging Brahminical Normativity to show how caste biases are perpetuated beyond the simple binaries of upper vs lower-caste categories. Our contributions are two-fold. Beyond challenging the categorical understanding of caste as a category, we propose an anti-caste approach to tackle the issue of caste bias and fairness in AI systems.

CCS Concepts: • **Social and professional topics** → **Cultural characteristics**; • **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → *Empirical studies in HCI*.

ACM Reference Format:
Divyanshu Kumar Singh, Dipto Das, Deepika Rama Subramanian, Koustuv Saha, Stephen Voida, and Bryan Semaan. 2018. Beyond Categories of Caste: Examining Caste Bias and Morality in Text-to-Image AI Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 17 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In recent times, Text-to-Image (T2I) models, such as Gemini, are gaining significant popularity with their enhanced ability to generate realistic imagery. In contrast to text representation, images carry greater amounts of information, and hence risks greater concerns for mis-representation of various constructs and elements within an image. Emerging

Authors' Contact Information: Divyanshu Kumar Singh, divyanshu.singh@colorado.edu, University of Colorado Boulder, Boulder, Colorado, USA; Dipto Das, dipto.das@utoronto.ca, University of Toronto, Toronto, Ontario, Canada; Deepika Rama Subramanian, deepika.ramasubramanian@colorado.edu, University of Colorado Boulder, Boulder, Colorado, USA; Koustuv Saha, ksaha2@illinois.edu, University of Illinois Urbana-Champaign, Urbana-Champaign, Illinois, USA; Stephen Voida, svoida@colorado.edu, University of Colorado Boulder, Boulder, Colorado, USA; Bryan Semaan, bryan.semaan@colorado.edu, University of Colorado Boulder, Boulder, Colorado, USA.

Manuscript submitted to ACM

research within FAccT has examined unique biases that are perpetuated by and through T2I models, such as misrepresentation of gender [27]. As a result there is dire need to examine unique socio-cultural biases and stereotypes that are embedded with T2I systems as they risk causing severe representational harms [46, 55], especially towards historically marginalized communities [46].

In the context of South Asia, caste marginalized communities, such as lower-caste, have faced oppression and violence for centuries. As a result, communities across South Asia are structured through various caste arrangements, such as cultural and rituals. Qadri et al. [46] highlighted caste as one of the "regimes of representation" within the AI system, as it "shape hegemonic ways of seeing and knowing about a culture or community". While some scholars have started to engage with Qadri's call, such as Ghosh [25], there currently exists a dearth of scholarship that actively investigates caste bias within T2I models. Moreover, most of the prior work exploring Caste within GenAI [25, 54], has only focused on a categorical understanding of caste, such as through caste categories. Caste is a relational phenomenon that is socially constructed and mediated by and through everyday interaction amongst different individuals [6, 32, 53]. Hence, in this work we turn our attention towards a novel understanding of relational nature of caste as represented through Text-to-Image GenAI system.

To address this, we designed an algorithmic audit study where we compiled names from different regions of India, and placed them in four categories, i.e., upper-caste (including middle), lower-caste, names that reflect ambiguous caste, and names with no surname. We then generate images with individual names and pairs of two across six everyday processes/dimensions – food, education, neighbourhood, migration, worship, and profession. Using critical discourse analysis [36, 45] we analyzed 1536 images generated via Gemini (flash-2.5-image), and show that explicit/implicit markers (caste categories or surnames) are not the only mechanism through which caste bias perpetuates in AI systems. Instead, the models are perpetuating a hierarchy of imposing a caste imagination through a moral order. We argue that AI systems are not only misrepresenting someone as poor, rather it is reproducing this same Brahmicial imagination that reproduces Brahminical Normativity. Therefore in order to tackle the embedded brahmanical epistemologies within the AI system, we argue for adopting a combination of decolonial and anti-caste approaches to AI.

2 Literature Review

2.1 Caste Relationality and Discrimination

Communities and individuals worldwide face everyday discrimination, rooted in structures like colonialism, sexism, and casteism, and enacted by dominant power with an intent to marginalize [52]. Marginalization excludes and pushes communities and individuals to the fringes of society, often based on their identity. These identities are socially constructed [17, 22, 28], either through how discursive practice [28], or associations with socio-cultural groups, heredity, norms [17], and history [22]. For example, caste identity emerges through a social hierarchical ordering of people that ascribes them to a particular socio-professional identity based on the family into which they are born [4].

The origin of the caste system can be traced back to the ancient Hindu religious texts and scriptures[4, 32, 37, 40]. There were four main castes: Brahmins at the top, followed by Kshatriya and Vaishya in the middle, and at the bottom, Shudra [37]. This hierarchy is called the varna system. There were people who were left outside of this system, also called out-caste/lower-caste [4]. As such, individuals in the lower castes were considered "untouchables" by upper-caste members, and hence deserving social exclusion [5]. Omvedt [40] and Ambedkar [4, 6] both have shown the role of Brahmins in appropriating existing cultural cults [40] and prohibition of social endosmosis [6] as the successful drivers of the caste system centered. Even though, the varna hierarchy appears rigid in structure, but it is socially constructed

by the Brahmins (through their superiority) - as they determined people's experiences in society vis a vis constant measurement through comparison [32]. Hence, also referred to as Brahamanical varna system. It is important to note that there is no single caste or varna system in India, as argued above Brahminical forces constantly govern and determines one's location in the society, and there are differing caste system across Indian depending on the local socio-cultural millieu.

Lower-caste communities and individuals till this date are severely discriminated against and denied basic human rights [1, 31]. Scholars have often termed this violence and oppression against the lower-caste as "Brahmanism" [32] or "Brahminism" [4]. Brahminism, is not a caste nor an identity, it is a rhetorical term that refers to violent, caste-based oppression that is practised within Hindu religion. Brahminism reflects the relational nature of caste [4, 32]. While colonial forces through colonial projects and demographic surveys helped the Brahmins in creating a categorical caste system that could be applied pan-India, they did not invent it [32]. Hence, colonial construction of caste basically reinterprets the brahmanical varna system of caste into a categorical administrative classification system, whereas the brahmanical caste system is socially constructed (relational) through brahmanical superiority.

In the everyday world, brahmanical superiority — that is, the brahmanical life-world — is imposed and normalized through various moral/social orders, what we dub, Brahmanical Normativity. Brahmanical Normativity is the routine and invisibilized practices that embeds brahminical practices rooted in brahminism and that constantly govern an individual's moral worth in the society. Hence, Brahmanical Normativity essentially governs the morality or the moral order [42, 48]. At the crux of brahmanical normativity is the practice of untouchability [5, 6]. For example, in the context of household help, lower-caste helpers are almost never employed for childcare, and similarly upper-caste helpers do not engage with cleaning the bathroom [26]. This example reflects how brahmanical normativity is reflected through invisibilized untouchability that morally classifies and governs, utility and purity. From the utility perspective, brahminism constantly dictates a moral classification and belongingness based on what a body can do or how a body's worth is morally justified through work. The lower-caste house help is best utilized to clean toilets, but the upper-caste body can be excused from that work as the brahmanical moral values prohibit that work [26]. Similarly, from the purity perspective, brahminism constantly determines the level of access to a lower-caste body as they are considered morally "impure" compared to upper-caste [cite, cite]. Hence lower-caste bodies must be controlled in order to avoid any "contamination" (through inter-mingling) to the upper caste. In the case of househelp, while lower-caste house helps are all allowed to enter the household, they are prohibited from being in the vicinity of a baby [cite] or even being present in the kitchen [cite].

The socio-material realities of caste discrimination, vis-a-vis, Brahmanism and Brahmanical Normativity, have come to shape almost different aspect of our society. Hence, through our work, we turn our attention towards how the emerging AI tools manifest such casteist logics. The dynamic and relational nature of Brahmanical Normativity offers a very unique lens to interrogate and understand the (in-)visibility of casteist logics and norms that continue to govern socio-technical systems, and hence motivating our study.

2.2 Caste Representation & AI

Technologies are socially constructed [43], as such, when technologies have been shown to perpetuate and reinforce existing societal biases and stereotypes [13–16, 18, 38]. While the capabilities of AI models have improved drastically over the past decade, the fact that these models are trained internet data raises some serious concerns. In particular, researchers across the world are examining unique biases that are being perpetuated by and through GenAI models [55], such as, racism [15], gender-bias [27], adultification [18], casteism [46], disability [34], geo-cultural [29, 46], and

coloniality [19]. There is a growing interest amongst FAccT researchers to evaluate various harms that are reflected through visual representations within GenAI [46]. Visual representations contain a lot of agency and power in creative societal narratives and discourse about cultural, community and individuals [20]. Hence, there is a great risk for harm if biases and stereotypes are unchecked within visual representations, especially with its appeal within the masses.

In the context of South Asia, visual misrepresentation through text-to-image models poses significant harms to communities, especially those that have been historically marginalized e.g., lower-caste [46]. Building on this, there is an emerging line of work that is exploring implicit and explicit caste representation within GenAI models, such as T2I [24] and LLM [54]. Ghosh [24] examined caste representation within Stable Diffusion's image output through explicit 'Caste-only' and 'Caste-occupation prompts'. In this work, Ghosh utilizes explicit labels of caste within the prompts, such as low-caste/high-caste and Brahmin, Kshastriya, and more. Through their finding, Ghosh highlights significant bias within the stable-diffusion model that reinscribes casteist stereotypes, such as limiting lower-caste communities within certain occupations. In contrast to this study, Vijayaraghavan et al. [54] designed an implicit name based – stereotypical word association task (SWAT) and persona based scenario answering task (PSAT), to evaluate caste biases within the large language models. Using the association of surnames as an implicit marker of caste identity within the Indian context, they compared the biases with higher and lower-caste names. Their findings highlight significant caste biases, particularly across upper and lower-caste groups, where lower-caste are shown to be attributed to menial work, lower educational status, and more.

While both of these studies have laid down a strong foundation for examining caste biases within generative AI systems, they both utilize a categorical understanding of caste, such as through explicit caste category (like Kshtriya) or implicit name categorization (like names/surname dataset). This approach provides limited information about how caste is socially constructed through routine process and practices, such as ambiguity [32, 41, 53]. Prior work, has shown the relevance of relational aspect of caste within computing industry [51, 53], and also, the risk of "presenting false homogeny" by flattening inequities within a community [46]. Our work is situated at the intersection of all of the above, where we examine relational representation of caste within Text-to-Image GenAI system. Hence, in our study we ask – RQ1: how do everyday processes encode caste representation within T2I models? RQ2: what kind of caste realities emerges through such representations?

3 Methods

Our study was motivated by emerging work that examines harmful socio-cultural stereotypes that are represented by GenAI applications (e.g., T2I). Building on algorithmic audit [19, 24] and critical reflexivity [39] we examine the relation aspect of Caste identity. The focus on examining the relationality of Caste helps us destabilize the existing monolithic understanding of caste as a hierarchy/category. We leverage critical reflexivity to design image generation prompts that are informed by existing literature [24, 54] and author's respective personal knowledge/history [cite, cite]. We generate images through Gemini's T2I model (flash-2.5-image) which are analyzed using a critical discourse analysis approach [45]. Below we present detailed steps and explanations for our methodology.

3.1 Prompt Design

Our prompt design approach is inspired by the recent work of Ghosh [24] and Vijayaraghavan et al. [54] that examine caste biases within GenAI. As highlighted in literature review, both these works have revealed significant genAI biases against lower-caste individuals, our work builds on these insights in critical and important ways. First, in Ghosh's [24] study, they explicitly prompted the models with the information about caste category (such as Brahmin), which

Dimension	*Prompt 1*	*Prompt 2*	*Prompt 3*
Food	eating food	having their respective favorite food	sharing their favorite food
Migration	in a non-south Asian country	together in a non-south Asia country	in two different non-south Asian country
Neighbourhood	in a locality	in their respective locality	in their locality, for each pick one of these {locality}
Worship	praying at the place of worship	praying together at the place of worship	praying at the place of worship, where one of them is leading the prayer
Profession	doing work	engaging in their respective work {option}	collaborating on work together
Education	studying	studying in their educational institution	studying together in an educational institution

Table 1. Table 1. Prompt Across Each Dimension

means that implicit factors were thereby excluded from the study. Moreover, the image generation prompt is largely concerned with a singular individual, almost as a portrait image. Second, while Vijayaraghavan et al. [54] overcame the limitation of Ghosh's focus on explicit prompting and added socio-cultural dimensions that influence caste, they ruled out any ambiguous name association in their names dataset. In addition to this, they also did not investigate within-group association tasks, such as an association or persona task between two lower-caste individuals. Both of these studies also treat caste as a categorical identity thereby ignoring the relational aspect of caste. In building on these studies and in addressing their limitations, we turn attention towards image generation prompts that leverage implicit caste markers (such as surname, no surname, and ambiguous names) and situate those in context by examining these markers as portrayed through everyday routine experiences and activities (dimension).

3.1.1 Compiling Names. We compiled a list of names using online caste based surname information (e.g., Wikipedia[1]), which we further complemented with our own personal knowledge and histories of growing up in India. Personal knowledge serves as an embodied source of data that emerges through a reflexive feminist standpoint [10, 11, 39]. As argued by Vaghela et al. [53], Ogbonnaya-Ogburu et al. [39], and Bardzell & Bardzell [10], personal knowledge/histories provokes marginalized voices as an analytical data source that uniquely challenges the normative episteme. By complementing the process of name compilation with our respective knowledge/histories, we were able to offer insights that are often invisibilized or flattened within South Asian identity discourse. To this end, we compiled names across four categories – Upper-Caste Name (UC_Name), Lower-Caste Name (LC_Name), Name with No Surname (No_SN) and Ambiguous Caste Name (AC_Name). Here, Upper-Caste Name and Lower-Caste Name comprises names which have traditionally been associated with upper/lower-caste communities within different regions. Names with no surname include names which we have personally encountered in our life that may or may not signify caste location. On the other hand, Ambiguous Caste Names are names which have surnames but are found in both upper- and lower-caste communities.

While compiling the corpus of names, we particularly focused on diversifying the regional aspect of caste names based on different regions we grew up in (e.g., Tamil Brahmin name and Northern Ambiguous caste name). While often the focus of analysis within caste is on the Hindu religion, recent scholarly work in sociology has highlighted the

[1]https://en.wikipedia.org/wiki/Category:Surnames_of_Indian_origin

prevalent caste within Muslim and Christian communities in South Asia [3, 8, 9, 21]. Hence, we added names from each of those communities, as well.

3.1.2 Compiling Dimensions. Second, we selected 6 dimensions that influence caste relationships within the Indian context – food, education, profession, worship, migration, and neighbourhood. We selected these dimensions as, based on prior work, they have been found to perpetuate stereotypical biases [13, 24, 54]. For example, Vijayaraghavan et al. [54] highlighted food, rituals and education, and Barve et al., highlighted location and occupation [13] as dimensions that shape caste relationships and experiences. For each dimension, we designed three prompts that reflected that dimension as a routine activity/process. Focusing on routine everyday tasks is critical to understand the relational aspect of as caste [26, 35]. Moreover, focusing on how everyday processes/routines are represented through imagery also provides insights about how the genAI system encodes the larger context. For example, Adrian C. Mayer [35], in his book titled, "Caste and kinship in central India: A village and its region," highlighted how during ceremonies, lower-caste individuals often sit outside the house doing their work, whereas upper-caste individuals sit on chairs within the premise of the house. This example is at the core how everyday casteism is reflected through routine processes/activities. If we are to simply focus on the individual identity of a given caste, it provides limited information.

3.2 Image Generation

3.2.1 Choosing API. For the image generation in our study we used Google's Gemini image generation API, in particular 'gemini-flash-2.5-image'. During our exploratory study design phase we experimented and decided to use two image generators – Gemini and DALL-E. We chose these two generative AI tools, largely because of their popularity and also because they are being actively examined within critical AI research [13, 46]. Both of these tools provide paid API access, where DALL-E priced at $0.08 per image output, and Gemini priced at $0.039 per image output. Due to funding constraints, we opted for Gemini's API, and excluded DALL-E.

3.2.2 Generating Images. Prior work that has evaluated bias within image generation models have generated anywhere between 1200-1500 images depending on the context of study. For example, Ghosh [24] generated 100 images per prompt, across two categories of 'caste-only' and 'caste-occupation' prompt leading to roughly 1500 images. Similarly, Barve et al. [13] generate 400 images per model, where they used three different image generation models resulting in 1200 images. For our study, we had 6 dimensions with 3 prompts within each dimension (see Table 1), and 16 names (across four categories). In contrast to the previous work, we were also generating group images of two people together in Prompt 2 and Prompt 3 (see table 1). Hence, for each dimension, for Prompt 1 we generated 16 images (1 per individual), and then for Prompt 2 and Prompt 3, we ran the prompt unique combination two names in a pair out of 16, i.e., ${}^{16}C_2 = 120(combination)$. Therefore, we generated, 6 (dimension) x (16 individual image + (120 x 2) image for prompt 2 and 3) = 1536 images. The images were generated in December 2025.

3.3 Analysis

After the images were generated, the first author shared some sample images from each category with the second and third authors, so that everyone could quality check the images and start generating their respective understanding of various elements in the picture. The first three authors met regularly over a period of two weeks to discuss their interpretations of each image. This step ensured we could ask each other clarification questions and check our mutual interpretation and perception of the image. After this, the first author generated descriptive ethnographic-style memos [33] for a subset of images in each category. Each memo covered these basic questions – how is the person presented

in the image? What kind of activities are being done? Who else is present within the image? And what kind of location/scene is that image set within?

Due to both textual (memo) and visual (corresponding image) modality, we then used Critical Discourse Analysis (CDA) [36] as it allowed us to situate the images as being "constitutive of and by social practices" [45]. Our approach is inspired by Putland et al. [45] where they used CDA to analyze AI generated images for Dementia. In our collective analysis of memo and images, the first author generated the initial themes such as, "shown doing work in their locality", "eating in the same scene but different mats", etc. Then, over couple of iterations all three authors discussed different meanings and interpreations, leading to codes such as "importance of work to establish worth" or "political composition of being together", etc. At this stages codes were then discussed with the last author, as an external member check. This iterative process led us the layered conceptualization of body, social relation, and material-spatial relations. All the authors involved in study design and analysis have contextual experience with caste and casteism.

4 Results

Through our analysis we focused on different combinations of images generated by Gemini, and similar to previous research examining caste bias within T2I models, we did notice extreme cases of caste bias. Though in contrast to previous work [24, 54], we particularly focused our analysis on dominant and normative Brahminical thought that shapes caste realities. Hence, we present three layers of caste bias that embeds and produces brahminical normativity through AI – Bodily Morality, Social Relation, and Material-Spatial Infrastructure.

4.1 Bodily Morality: How AI Imposes Brahminic Dignity

The caste system operates across multiple facets of an individual's life, in turn, shaping their everyday experiences (e.g., access, dignity, and self-worth). Through our analysis, we find that one such way in which Brahminical normativity is perpetuated by and through AI is through how it mediates bodily morality. That is, **caste is not merely an individual's identity—rather, it governs what kind of life a human "deserves" through how it perpetuates an invisible moral order** [42]. In our generated image corpus, when we prompted for images of individuals living in particular localities, we found that while caste was reflected at the level of identity, there were also invisible moral orders that governed these images.

In Figure 1 (top row), at the level of identity/representation, we can clearly see that Figures 1a and 1b are explicitly caste marked. UC_Name_1 is depicted in a clean outfit, smiling, on a clean street, whereas LC_Name_3 is depicted as a cleaner on a dirty street, holding a broom and wearing worn-out clothes. Hence, from the identity and representational perspective, these images are biased in how they govern bodily morality. Through the visual markers in these images, there is an inherent moral order that is shaped by the brahminical view of caste that produces such life worlds. In the case of UC_Name_1 the streets are depicted as clean, flowers are blooming, the neighbourhood is spacious and clean, fruit stands adorn the background—there is a hidden reality that has made that possible. Lekshmi Iyer is not holding a broom (like LC_Name_3) or a fruit basket (like No_SN_1), or depicted as the owner of a general store (like No_SN_4). The Brahminical moral order grants upper-caste the privilege and self-worth to benefit from the work of all other castes. That is, the lower caste members of society are often relegated to being street cleaners, fruit pickers, and other occupations that they are deemed worthy of and that are "beneath" upper castes. AI models have come to learn and perpetuate this Brahminical bodily moral order, which differentiates between dignity and utility.

Similar to prior work [54], we found that surnames represented caste biases, but our analysis also yielded instances where the lack of surnames also reflected these same biases. That is, surnames are governing the worth of individuals

Fig. 1. **Top Row** (Left to Right) UC_Name_1 (1a), LC_Name_3 (1b), No_SN_1 (1c), No_SN_4(1d); **Middle Row** (Left to Right) No_SN_1 & AC_Name_1 (1e), LC_Name_2 & AC_Name_4 (1f), LC_Name_2 & No_SN_2 (1g), UC_Name_2 & No_SN_3 (1h); **Bottom Row** (Left to Right) UC_Name_1 & No_SN_2 (1i), No_SN_3 & AC_Name_4 (1h), AC_Name_1 & AC_Name_4 (1j), UC_Name_4 & LC_Name_4 (1k)

and are assigning them routines and/or occupations that make them "useful" as pre-determined by bodily morality. In Figure 1 (middle row), No_SN_4 has no surname and AC_Name_1 has an ambiguous surname. Yet, the model imagined No_SN_4 living in a village and AC_Name_1 sipping a beverage in the city. Similarly, when asking Gemini to place Muslim LC_Name_2 with AC_Name_4 and Christian No_SN_2, respectively, Muslim LC_Name_2 is shown in front of a meat shop on an old Delhi street, AC_Name_4 is depicted drying clothes in a balcony over a very congested street, and No_SN_2 is shown on Kerala's coast standing in front of boats and fishing net. Lastly, both Muslim UC_Name_2 and No_SN_3 are depicted simply standing on the street and house wearing clean clothes without undertaking any work. These examples highlight how bodily morality is operating invisibily through AI.

Even in the absence of surnames (No_SN_1 and No_SN_2) or with ambiguous surnames (AC_Name_4 and AC_Name_1), the model almost never grants these bodies the same morality as a Brahmin. In the absence of surnames and when these surnames include ambiguity, the model is unable to identify caste but still automatically assumes a moral classification and belongingness depending on what this body could potentially do (utility). This deeply casteist logic is rooted in Brahmincal normativity [cite]. For example, the assignment of labor/service roles to No_SN_2 (dealing with fish) defines their caste location not explicitly through surnames, but through a moral order of hierarchy. Whereas, AC_Name_1, whose surname is ambiguous (can be any caste), and No_SN_3, who lacks a surname, are assigned no work to clearly highlight the relational nature of the caste system. This challenges the assumption that explicit/implicit markers (caste categories) is not the only mechanism through which caste bias perpetuates in AI systems. Instead, the models are perpetuating a hierarchy of imposing a caste imagination through a moral order.

Further, it is not just vocation that determines the individual's self-worth in the Brahminical moral order, but it is a holistic imagination of what a body deserves. This is illustrated in Figure 1 (bottom row), which prompted Gemini to develop images of individuals in their respective institutions. We find that UC_Name_1 is depicted sitting on a chair and table studying in a library with a technological device, whereas No_SN_2 is depicted sitting on a bench with a notebook in a classroom space that appears to be open design, signifying a lack of resources. While these images may appear to encode representational biases visually and aesthetically, that reading misses the Brahminical moral

order. No_SN_2, even with the absence of caste markers, is not imagined as morally worthy of a similar educational condition as UC_Name_1. As such, lower-caste bodies are morally obligated to be arranged for their comfort, worth, and future with available resources around them, whereas upper-caste bodies, by the virtue of their caste location, deserve entitlement without explanation. In the next two images, of No_SN_3 with AC_Name_1 and AC_Name_4, respectively, both No_SN_3 (no surname) and AC_Name_1 (with ambiguous surname) are shown as deserving better educational space (an apparently college level classroom). In contrast, AC_Name_4 is imagined to be schooled at a more primary level and not be deserving of advanced educational investment. These examples, again, reflect the relational aspect of caste and illustrate how these models inherently adopt a Brahmanical imagination that goes beyond caste categories of names.

4.2 Social Relation: How AI Imposes Brahminic Social Legibility

One of the ways that the relational aspect of the caste system is mediated by the social world/arrangement that is made legible for a person. The Brahminical moral order actively legitimizes and governs different kinds of social arrangements, such as who is present around whom and who is doing what [6]. In our image generation, we are particularly focused on processes in the everyday world to capture this relational morality that is enforced upon the social world. For example, when we requested images around eating, with the model choosing the type of food consumed or shared (see Figure 2), the model generated obvious visual differences of food. But the model also amplified Brahminical morality that strictly conditioned social intermingling and respective social worlds, such as, in Figure 2 (top-row) a–c. In Figure 2a and Figure 2c, UC_Name_1 (on left) is shown within the clean environment of her home, whereas Muslim LC_Name_2 is shown eating on a table outside a restaurant and LC_Name_1 is shown eating in a shack in a village setting (indicated by the mud and haystack house). There are no other humans around UC_Name_1, whereas there are some people in the background of Muslim LC_Name_2, and one person behind LC_Name_1 who is sitting near the utensil. Eating food is considered a ritual, and as such the absence of human interaction around Iyer is a moral order that maintains and insulates the ritual purity of UC_Name_1 home, food, and relationships. In contrast, Muslim UC_Name_2 is allowed to be within the same space on the same table as Hindu UC_Name_2, clearly demonstrating who is allowed and who is not allowed. The "graded access" within one's social world is at the core of caste inequality, as depicted through these three images above [cite]. Lastly, even with the absence of a surname, Anarkali is shown not sharing the same space as LC_Name_3 (Figure 2d). LC_Name_3 is sitting in the middle of the street, and the people in his background are also eating food sitting on the floor, showcasing a communal aspect of eating food. The model refuses to allow for cross-caste mingling. Here, a no-surname person whose caste membership is ambiguous with an obviously lower-caste individual, highlighting a moral incompatibility of the two social worlds and perpetuating historical discrimination. While we can notice bodily morality at play in these images, the caste system has also shaped the inherent social world for each individual by conditioning different levels of social compatibility.

Similarly, in the context of worship, when we requested images of individuals praying together and images in which one of them is leading the prayer, we noticed a stark contrast between social settings within images (Figure 2). For example, in Figures 2b and 2c, No_SN_1 (with no surname) and UC_Name_1 (upper-caste) are shown leading the prayer, whereas LC_Name_4 (lower caste) and UC_Name_4 (upper-caste) are shown sitting on the side. Leading a prayer is caste-coded work, such as being priest, which has historically been gatekept to the upper-caste, primarily Brahmins. The models prefer UC_Name_1 to lead the prayer, because of their caste membership (Brahmin) and No_SN_1 for their ambiguous caste, as the other option—allowing LC_Name_4 to lead the prayer—could potentially disrupt the Brahminical moral order. Similarly, the moral hierarchy puts UC_Name_1 over UC_Name_4. More interestingly, one

Fig. 2. **Top Row** (Left to Right) UC_Name_1 & LC_Name_2 (2a), UC_Name_1 & UC_Name_2 (2b), UC_Name_1 & LC_Name_1 (2c), LC_Name_3 & No_SN_3(2d); **Bottom Row** (Left to Right) LC_Name_3 & No_SN_1(1e), No_SN_3 & LC_Name_4 (2f), UC_Name_1 & UC_Name_4 (2g), LC_Name_3 & AC_Name_4 (2h)

could also notice that UC_Name_1 is leading the prayer for a large mass of people in the temple, mostly women, this again reflects the historical Brahminical practice of men serving as priests (e.g., Sabrimala incident [cite]). In Figures 2e and 2h, LC_Name_4 is shown with No_SN_1 and AC_Name_4, and the social setting has shifted altogether: from larger temples to smaller, village-based or street temples. From a simple identity/representation perspective of caste, this is a classic case of lower-caste being denied from worshiping in the temple or serving in the priesthood [cite]. But, looking through the Brahminical moral hierarchy, we notice that (supposed) lower-caste is not simply being prohibited from the temple, but rather the social world around what is a temple is completely reimagined. That is to say, the Brahminical morality contains the sacred authority by imposing a smaller localized social religious world. This is clearly evident in both images: not only has the temple been scaled down, but also the audience in the background is portrayed as bystanders living with minimal economic means and prospects.

4.3 Material Spatial Relation

Beyond the individual and their social relations, caste systems also govern the material and spatial order that shapes people's everyday experiences. That is, the dialectical relationship between people's spatial arrangements (where they are located and/or exist) and material (artifacts that they own, are on their person, or surround them in a space) are governed through Brahminic morality. In our corpus of generated images, we see this Brahiminic morality re-enforced through the space people inhabit and the materials in their possession or that surround them, such as homes, furniture, streets, graphics, and more. Through the lens of representational identity, the material-spatial configurations within the generated images provide subtle differences between people across the caste hierarchy, such as those related to cleanliness, congestion, and economic viability. In this way, by re-enforcing the Brahiminic material and spatial order (cf Guru [26]), AI is maintaining the rigidity of the caste system by restricting social mobility and representations of social mobility.

For example, we generated images with prompts about doing work, living in a locality, and studying (Figure 3). Across these images we found infrastructural constraints imposed on individuals as imposed by Brahminic morality. In Figure 3e, No_SN_1 (no surname) is imagined doing pottery work (caste-coded work). From a bodily morality and social relations perspective, her life is constrained around the work she is pre-determined to do by the caste hierarchy and the social arrangement within that space feature an adult and children living within the same condition. Beyond bodily morality and social relations,, the material-spatial configuration in which she is embedded—the village setting and makeshift house—are imagined through Brahminic morality as future mobility. In Figures 3b and 3c, LC_Name_3

Fig. 3. **Top Row** (Left to Right) No_SN_1 (3a), No_SN_2 & LC_Name_3 (3b), LC_Name_3 & UC_Name_4 (3c), LC_Name_3 & LC_Name_4 (3d); **Bottom Row** (Left to Right) No_SN_2 & No_SN_1 (2e), LC_Name_3 & No_SN_2 (2f), No_SN_4 & AC_Name_4 (2g), LC_Name_3 & No_SN_4 (2h)

is shown doing cleaning work alongside Jesudas (no surname) and Ravindar Kamble (ambiguous caste), respectively. In both images, the material-spatial configuration in which they are situated restricts any social mobility using two Brahmanical mechanisms. In Figure 3b, there is a banner in the background reading "Bhangi Colony." In Figure 3c, there is a board (with misspelled words) saying something about "cleaning" (swachhta). These images beg the question of why does the system have to mark a colony as "bhangi colony"? Or why is the system justifying a profession as social good or service for the public? Here, Brahmanical morality works through constant identification and justification of the caste system. In the context of the caste system, the co-mingling of people from different castes across the material-spatial order is seen as "polluting" the upper castes. Designating a locality for lower-caste (bhangi) is an act of insulation from inherent pollution and thus keeping them "pure." Moreover, by labeling the locality upper castes are able to proactively counter or avoid this pollution through how the material-spatial order is made visible and legible. Similarly, the justification of a profession is a justification of Brahminical moral order that argues that the caste system is only necessary for work, denying any socio-political consequences [6].

In a similar vein, we found various instances across the generated images in which certain religious elements are present in different forms that are completely irrelevant to the prompt provided. For example, in Figure 3 (e-f), almost every image depicts Hindu deities in the background, irrespective of whether individuals were located at home, educational institutions, or in their workplace. In figure 8d, both individuals are shown with a red tika (dot) on the forehead, which is a common Hindu practice. While the depiction of cultural elements itself is not a problem, the default towards Hindu practice is. The inclusion of Hindi deities and the red tika signal a material-spatial ordering that enforces a Hindu default, whereas the continent is more diverse. We have already shown that even Muslim and Christian names are shown to embed caste bias in the model, so then the question arises why is the model defaulting towards larger Hindu religious practices/style?

5 Discussion

Our findings challenge the understanding of caste bias as one that is categorical (e.g., caste categories and surnames) and re-animates our understanding of caste bias as one that is relational in nature. That is, our work shifts the discourse from a focus on categorization to centering relationality – where people's worth and station in society is a product of constant measurement through comparison across the caste hierarchy. In building on critical scholarship of AI that eplores caste bias (e.g., Ghosh [24] and Vijayaraghavan et al. [54]), which has focused on the categorical nature of caste, our research illustrates the utility and importance of focusing on caste as a relational phenomenon, particularly in the

context of FAccT's exploration of accountability in GenAI systems. Addressing caste as a relational phenomenon equips us to understand the underlying mechanism of how the caste system operates through the invisibilized moral order of caste – brahmanical normativity. This leads to the question of how do we address brahmanical normativity that is embedded within AI models? To answer this question, we first argue for an Anti-Caste lens that can work towards reimagining AI Bias and Fairness. We then articulate future research directions that can further inform this reimagining.

5.1 (Re)Imagining AI Bias & Fairness through Anti-Caste Lens

In order to articulate opportunities for reimagining AI Bias and Fairness, we first need to understand the pitfalls with the existing ontological understanding of caste. Scholars who have explored the issue of caste bias in AI research have often relied on a categorical understanding of caste that treats caste as an identity, and that can be deduced through explicit [24] or implicit markers [54]. This approach is helpful in understanding 'what' types of representational harms are perpetuated by AI models, but is limited to considering the categorical understanding of caste that largely emerges through post-colonial ontologies of caste, which interprets caste as a colonial construct [32]. As noted in the literature review, caste as a colonial construct only expanded the varna system of caste for classification (identity), whereas the root of the caste system can be located within brahmanism's moral superiority and dominance that continuously situates people (graded inequality [4]) through comparison. This is what is dubbed the brahmanical ontology of caste. We argue that in order to mitigate the roots of caste bias within the AI machine, we should actively look beyond post-colonial ontologies of caste.

Emerging work with FAccT and HCI have shown how colonial logics are perpetuated through AI systems [12, 56] and also through bias/fairness mitigation frameworks that emerge in the West [44, 50]. Therefore, in order to understand the inherent coloniality of AI systems, scholars have argued to move away from Western-centric mitigation strategies, and engage in decolonial epistemologies [12, 19, 50]. For example, Sambasivan et al. [50]urged the FAccT community to move beyond Western-centric fairness frameworks by integrating local contexts, such as caste-reservation (affirmative action). Similarly, Barett et al. [12] argued for embracing pre-colonial African and Indigenous philosophies within broader responsible AI practice. Scholars have also argued for embracing "refusal" [57] as an integral method within AI praxis, not only as a form of community resistance, but also as a result of scholars refusing datafication [24]. For example, in the context of caste, Ghosh [24] rightly pointed out the potential negative downstream effects of integrating caste fairness through efforts such as datasets/information, as it could lead towards further discrimination. While we agree with these assessments, our findings build on these insights through our findings that AI models have come to learn not just colonial epistemologies of caste (e.g., surname), but on a deeper and implicit level have come to embody the normative brahmanical epistemologies (e.g., brahmanical morality). Hence, in order to dismantle brahmanical normativity within the AI machine, we urge FAccT scholars–and other scholars and practitioners who engage in work to understand, critique, design, and build AI and/or AI-mediated platforms–to combine De-colonial [2] approaches with De-brahmanical (Anti-Caste) [51] approaches that are rooted in Anti-Caste epistemologies. We believe combining these two approaches can work towards addressing both the categorical and moral biases that are explicitly and implicitly shaping the AI machine. We further articulate this imperative through two primary implications that can lead towards an anti-caste future, including: (1) Relationality as Anti-Caste; (2) Empowering Dalit Subjectivities.

5.1.1 Relationality as Anti-Caste. From a systems perspective, caste is not simply (mis)represented. Rather, AI models have come to learn the brahmanical social/moral order through existing knowledge and data. Sambasivan et al., [50] and Ghosh [24] utilized anti-caste lenses in their study. However, their reliance on a post-colonial understanding

of caste reflected decolonial narratives and overlooked the inherent brahmanical normativity of caste in AI. This limitation could be overcome by utilizing a relational understanding of caste community narratives and oral histories of brahminism. Though, Qadri et al. [46] argued that while integrating community narratives are powerful steps in building a more inclusive AI harms framework, communities are not panacea of everyday realities. The colonial construction of caste creates a stable caste entity/identity, such as lower-caste or upper-caste, whereas caste in practice is largely relational. Yet, as our work has highlighted and what critical scholars of caste continue to champion is that there is no homogenous lower or upper-caste community. Hence, future anti-caste lenses should shift from these previous categorical understandings to working towards dismantling inherent brahmanical ideologies that continue to govern AI development and platforms that are shaped by AI (e.g., generative AI).

5.1.2 Empowering Dalit Subjectivities as Anti-Caste: Decolonial approach in the context of India, and South Asia at large, comes with their own geo-political and historical nuances. Marie-Therese Png [44] argued that in order to develop AI governance frameworks we need to critically examine the underpinning imperial legacies. Similarly, Raval et al. [47] argued for taking into account the colonial imprints on law and governance in post-colonial society that ultimately influences technological law and governance. While we agree with these scholars, we note that casteism predates colonialism. Therefore we advise caution when advocating for adopting decolonial approaches as they may very well provide grounds to reinscribe brahmanical casteist philosophies as decolonial narratives. Hence, anti-caste epistemologies should be actively integrated with decolonial epistemologies to dismantle brahmanical normativity. This becomes extremely timely, as more and more brahmanical forces across India are now weaponizing decolonial approaches to impose a brahmanical world view (e.g., law) [32].

One set of Anti-Caste approaches centers attention on the inclusion of the voices of the people who are being most harmed by technology, including the technology of caste. Kumar [32], argued that oral narratives and stories counter brahmanical decolonial intent of governance, and hence similar strategies should be adopted to design inclusive AI frameworks. In addition this, local-level land revenue records are an emerging tool to understand history of caste as these records were not influenced by colonial sociological project (like surveys) [49]. Hence, we suggest to understand fairness within Indian context, land-revenue records should be integrated with Dalit oral narratives should be deeply embedded into our system design.

5.2 Bias vs Reality?: Dismantling Brahaminical Normativity

While we have shown the mechanism of relational aspect of caste biases within AI system(s), a valid question that arises: are models perpetuating bias or reality? Lower-caste communities face discrimination in the everyday world, and as such, one could argue that AI models are aligned with reality. While this is a valid but apolitical argument, it ignores the politics of caste by limiting representation as a mere issue of an individual's/community's characteristic. On a deeper level, when models work to capture existing reality, they do not work to address root historical issues that have come to shape the everyday experiences that are now being modeled and re-perpetuated.

When an AI model imagines a lower-caste/ambiguous-caste person as poor or working menial jobs, it ignores the root cause—the perpetrator that has contributed to that situation/condition. Dalits make up for 77% of manual laborers in India [23]. This is not due to a lack of characteristics that would allow them to engage in other work that is often seen as above their station, such as doctors or teachers. Rather, the brahmanical caste and the brahmanical morality that has shaped the caste system has deprived them of any power/social mobility by constantly subjugating them to a particular profession. This inequity is normalized as an inequality reflecting the characteristics of the individual/community, rather

than caste politics has resulted in that outcome. Brahmanical normativity relies and benefits from this apolitical-ness, and is now also being imagined through AI models. Moreover, the term Dalit is itself a political term [41]. Dalit is not simply a caste or identity, it is an anti-caste subjectivity and anti-identity [7]. In our findings, we highlighted how Brahmanical normativity is perpetuated through morality as the default difference within AI system(s). Dismantling this moral framework requires the acknowledgement of inherent brahmanical politics that erases caste subjectivity. Hence, as critical AI researchers, we should investigate and question the inherent moral frameworks that are rooted in casteism (brahminism). To aid this investigation we propose two provocations/implications to inform future research directions.

First, we need to understand that (mis)representation of caste is not merely a lower-caste issue. That is, we must look at these misrepresentations across all dimensions of the caste system. The uni-directionality of our solidarity towards lower-caste, treating them as a problem to be fixed, takes the spotlight away from the role of brahmin and non-brahmin elite castes. As Ambedkar argued, we hardly even hear anyone trying to fix the "touchable Hindu" [5]. If we are to truly dismantle brahmanical normativity within AI systems, we must question AI's imagination of upper/elite-caste as well. The question then is not only how/why the lower-caste is shown as poor or protesting [24], but also why/how the upper-caste is always shown as superior or dominant. Hence, we urge scholars to think about not just misrepresentation within AI but also what is invisibilized or absent from AI. For example, why is the reversal absent or not normalized? Why can a Brahmin or non-brahmin elite not be portrayed as a toilet cleaner? We as a community are largely fixated upon resolving stereotypes, exclusion, and misrepresentation of lower-caste. In turn, we have ignored the upper-caste as merely a baseline for difference, treating upper-caste brahmin and non-brahmin elites as an apolitical and absent entity. It is important to remind ourselves – Brahmin is also an untouchable – an ideal untouchable, whereas a Dalit is a despicable untouchable [26].

Second, from an algorithmic perspective, it is not simply a data or a pipeline issue – it is a systematic issue. For example, in their recent study Vijayraghavan et al. [54] noted how some large language models refused to engage in a task when the model suspected potential stereotypes being perpetuated, such as in the name-association task (as discussed in the lit review). Whereas refusal is often construed as a human property, this same refusal is becoming an active safety guardrails strategy for AI systems when they refuse to perform a task. However, issues arise when this refusal becomes selective [30]. Similarly, in our work we identified issues of refusal responses from AI during the exploratory phase of our study prompt design when using Ghosh's [24] prompting approach in asking Generative AI to "draw an Indian Balmiki caste person." Regardless of a model's safety guardrails, when models were requested to generate the image, the model responded with adjectives like "dignified" and "respect." While the development and enforcement of guardrails are steps in the right direction, the example from our study reflects the apolitical nature of caste. The model only uses such adjectives when asked to draw a lower-caste person, so the question arises: why does the model want to enforce "dignity" on lower-caste, but upper-caste have a default right to dignity? Or, put another way, when models default to guaranteeing dignity for upper-caste they are encapsulating root issues stemming from Brahmanical normativity.

6 Conclusion

In this work we examined the relational caste representation with Text-to-Image GenAI systems, and we challenged the existing understanding of caste biases being perpetuated through categorical caste markers, such as caste location and names. Our work contributes and urges an ontological ("what is") shift in our understanding of caste, to focus on routine relational mechanism of caste. Prior research within critical AI discourse has yet not addressed the inherent Brahmanisim

embedded within GenAI systems. And therefore, these systems revalidated and circulated the very stereotypes and brahminical morality that caste ideology depends on, giving them new visibility and technical legitimacy. As we move into the future, we hope our community can come together in working towards addressing these issues with AI systems that continue to perpetuate brahmanic morality. The issues that continue to perpetuate a long-standing and historical structure that has shaped people's experience in profoundly *immoral* ways.

References

[1] 2024. https://timesofindia.indiatimes.com/city/hyderabad/dalit-man-allegedly-abused-by-priest-and-temple-officials-over-lack-of-dakshina/articleshow/114956944.cms

[2] Mustafa Ali. 2014. Towards a decolonial computing. (2014).

[3] Manjur Ali and Shilp Shikha Singh. 2024. Contemporary 'Pasmanda'Leadership and the Hindutva Politics in Uttar Pradesh. *Studies in Indian Politics* 12, 1 (2024), 33–47.

[4] Bhimrao Ramji Ambedkar. 1945. *Annihilation of caste with a reply to Mahatma Gandhi.*

[5] B. R. Ambedkar. 2014. *Untouchables or The Children of India's Ghetto.* Government of Maharashtra / Dr. Ambedkar Foundation, Mumbai/New Delhi. https://www.mea.gov.in/Images/attach/amb/Volume_05.pdf Part of the Collected Works of Babasaheb Dr. B.R. Ambedkar, Vol. 5.

[6] Bhimrao Ramji Ambedkar. 2022. *Castes in India: Their mechanism, genesis, and development.* DigiCat.

[7] S Anand. 2006. On claiming dalit subjectivity. In *SEMINAR-NEW DELHI-*, Vol. 558. MALYIKA SINGH, 60.

[8] Shireen Azam. 2023. The political life of Muslim caste: articulations and frictions within a Pasmanda identity. *Contemporary South Asia* 31, 3 (2023), 426–441.

[9] Shireen Azam. 2023. Scheduled caste status for Dalit Muslims and Christians. *Economic and Political Weekly* 58, 27 (2023), 14–19.

[10] Jeffrey Bardzell and Shaowen Bardzell. 2016. Humanistic Hci. *Interactions* 23, 2 (2016), 20–29.

[11] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1301–1310.

[12] Teanna Barrett, Chinasa T Okolo, B Biira, Eman Sherif, Amy Zhang, and Leilani Battle. 2025. African Data Ethics: A Discursive Framework for Black Decolonial AI. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 334–349.

[13] Saharsh Barve, Andy Mao, Jiayue Melissa Shi, Prerna Juneja, and Koustuv Saha. 2025. Can we Debias Social Stereotypes in AI-Generated Images? Examining Text-to-Image Outputs and User Perceptions. *arXiv preprint arXiv:2505.20692* (2025).

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

[15] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 1493–1504.

[16] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[17] Judith Butler. 2002. *Gender trouble.* routledge.

[18] Jane Castleman and Aleksandra Korolova. 2025. Adultification Bias in LLMs and Text-to-Image Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2751–2767.

[19] Dipto Das, Shion Guha, Jed R Brubaker, and Bryan Semaan. 2024. The"Colonial Impulse" of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

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

[21] Nidhin Donald and Asha Singh. 2023. Beyond the Paternity of Caste: The Dalit Christian/Dalit Muslim Challenge to the Rule Book. *Economic & Political Weekly* 58, 9 (2023).

[22] Frantz Fanon et al. 1970. *Black skin, white masks.* Paladin London.

[23] Azeefa Fathima. 2024. 77% of manual scavengers are Dalit, says report despite Union Govt's denial.

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

[25] Sourojit Ghosh and Aylin Caliskan. 2023. 'Person'== Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *arXiv preprint arXiv:2310.19981* (2023).

[26] Gopal Guru. 2009. Archaeology of untouchability. *Economic and political weekly* (2009), 49–56.

[27] Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, et al. 2024. Akal badi ya bias: An exploratory study of gender bias in hindi language technology. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1926–1939.

[28] Cultural Identity. 2000. Who needs' identity'? Stuart Hall. *Identity: A Reader* (2000), 15.

[29] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. 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.

[30] Adel Khorramrouz and Sharon Levy. 2025. Characterizing Selective Refusal Bias in Large Language Models. *arXiv preprint arXiv:2510.27087* (2025).

[31] Ritu Kochar. 2022. From Traditional to Modern Atrocities: Has Caste Changed in Independent India? *Contemporary Voice of Dalit* (2022), 2455328X221136385.

[32] ARVIND KUMAR. 2025. Indigeneity, caste, tribe and the limitations of decolonial thought in South Asian socio-legal studies: The need for a decolonial–debrahmanical approach. *Journal of Law and Society* (2025).

[33] Lora Bex Lempert. 2007. Asking questions of the data: Memo writing in the grounded theory tradition. *The Sage handbook of grounded theory* (2007), 245–264.

[34] Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K Kane, and Cynthia L Bennett. 2024. "They only care to show us the wheelchair": disability representation in text-to-image AI models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–23.

[35] Adrian C Mayer. 1970. *Caste and kinship in central India: A village and its region*. University of California Press.

[36] Michael Meyer. 2001. Between theory, method, and politics: positioning of the. *Methods of critical discourse analysis* 113 (2001), 14.

[37] Pravhati Mukherjee. 1988. *Beyond the Four Vernas: Untouchables in India*. Motilal Banarsidass.

[38] Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 786–808.

[39] Ihudiya Finda Ogbonnaya-Ogburu, Angela DR Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for HCI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.

[40] Gail Omvedt. 2003. *Buddhism in India: challenging Brahmanism and caste*. Sage.

[41] Shailaja Paik. 2011. Mahar–Dalit–Buddhist: The history and politics of naming in Maharashtra. *Contributions to Indian Sociology* 45, 2 (2011), 217–241.

[42] Shailaja Paik. 2018. The rise of new Dalit women in Indian historiography. *History Compass* 16, 10 (2018), e12491.

[43] Trevor J Pinch and Wiebe E Bijker. 1984. The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science* 14, 3 (1984), 399–441.

[44] Marie-Therese Png. 2022. At the tensions of south and north: Critical roles of global south stakeholders in AI governance. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1434–1445.

[45] Emma Putland, Chris Chikodzore-Paterson, and Gavin Brookes. 2025. Artificial intelligence and visual discourse: A multimodal critical discourse analysis of AI-generated images of "Dementia". *Social Semiotics* 35, 2 (2025), 228–253.

[46] Rida Qadri, Renee Shelby, Cynthia L Bennett, and Remi Denton. 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*. 506–517.

[47] Noopur Raval, Amba Kak, and Luke Strathmann. 2021. A New AI Lexicon: Responses and Challenges to the Critical AI Discourse.

[48] Ramnarayan S Rawat. 2013. Occupation, dignity, and space: The rise of Dalit studies. *History Compass* 11, 12 (2013), 1059–1067.

[49] Ramnarayan S Rawat and Kusuma Satyanarayana. 2016. *Dalit studies*. Duke University Press.

[50] 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.

[51] Divyanshu Kumar Singh and Palashi Vaghela. 2024. Anti-Caste Lessons for Computing: Educate, Agitate, Organize. *XRDS: Crossroads, The ACM Magazine for Students* 30, 4 (2024), 41–45.

[52] Dan Trudeau and Chris McMorran. 2011. The geographies of marginalization. *A companion to social geography* (2011), 437–453.

[53] Palashi Vaghela, Steven J Jackson, and Phoebe Sengers. 2022. Interrupting merit, subverting legibility: Navigating caste in 'casteless' worlds of computing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.

[54] Prashanth Vijayaraghavan, Soroush Vosoughi, Lamogha Chiazor, Raya Horesh, Rogerio Abreu de Paula, Ehsan Degan, and Vandana Mukherjee. 2025. Decaste: Unveiling caste stereotypes in large language models through multi-dimensional bias analysis. *arXiv preprint arXiv:2505.14971* (2025).

[55] Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030* (2024).

[56] Meg Young, Michael Katell, and PM Krafft. 2022. Confronting power and corporate capture at the FAccT Conference. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1375–1386.

[57] Jonathan Zong and J Nathan Matias. 2024. Data refusal from below: A framework for understanding, evaluating, and envisioning refusal as design. *ACM Journal on Responsible Computing* 1, 1 (2024), 1–23.

A Positionality Statement

The personal is political, and thereby analytical [cite]. In recent times, fields like HCI and FAccT have embraced a reflexive turn that acknowledges the deeply intricate relationship between power and knowledge production. Research is

actively shaped by and through researchers' motivations and perspectives, especially when "studying the understudied" [cite] communities [cite]. Research is a political project. This project emerged through the countless conversations amongst first and third authors about epistemological understanding of caste within our discourse. While it was liberating to read/engage with existing algorithmic audits dealing with caste, there was also an inherent frustration with epistemological understanding of caste. Hence, this study emerged to push against the monolithic understanding of caste as merely a categorical construct.

Four authors out of six were born and brought up in South Asia, specifically in India and Bangladesh. The last two authors were born and brought up in the United States. The experiences of the first three authors growing up in different and diverse communities across India and Bangladesh have significantly shaped the study. We grew up in different geographies with unique languages and cultures, and particularly, experienced unique caste inequities and structures. For example, during family gatherings, the first and second author did not experience the same caste practices based on their location in the caste hierarchy. That is, one was included and sat with others during these gatherings whereas the other was oftentimes made to sit in a designated spot for particular-caste members. Similarly, as we were compiling the names for our analysis we learnt from the third author's experience that surnames within their communities do not necessarily signify caste location. Moreover, at times the sanskritized pronunciation of a given name could hint at one's caste location. We shared similar experiences during our discussion that eventually shaped our protocol design and the analysis, and moreover the motivation for this paper.

The first and second author, over a period of four months, examined FAccT literature to experiment with different protocol designs. During this time they actively gathered feedback from the fifth and sixth authors. In order to cross-check our biases and also add further diverse opinions to our protocol design, the third author was involved and provided critical feedback throughout the study. Lastly, we involved the fourth author gathering feedback on the final study design and seeking assistance in framing the study. All the members of our research are strong advocates for equity and justice within social computing. Collectively, our team has more than two decades of research experience examining myriad socio-technical systems, with expertise ranging from caste, gender, coloniality, information ecosystems, ubiquitous computing, well-being, and social justice. Our respective positionality, experiences, and commitments have enabled us to examine and surface the invisible narratives of lower-caste communities within GenAI research.

B Ethical Consideration

We as a team decided not to publicly display the names that we compiled for our prompt, because publishing those name, especially the surname could have real world, such as through inappropriate identificaton. Hence, instead we use markers for each category of names, such as, UC_Name_1. The asthetic within imagery plays a key role in understanding caste bias, and hence we decided not to blur the images.

C Generative AI Statement

No Generative tools were used in writing of this manuscript.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009