

What Should AI Generate? Moving Beyond Bias in Generative Systems

Jennifer Chien

The McCoy Family Center for Ethics in Society and The Stanford Institute for Human-Centered AI (HAI)
Palo Alto, California, USA
drchien@stanford.edu

Abstract

Generative AI systems seamlessly transform underspecified prompts into embodied outputs, masking the numerous decisions required to instantiate a representation. These decisions not only shape what is depicted, but render implicit statements about what is possible, likely, and necessary. Existing work has largely evaluated such patterns through notions of representational bias. However, conflicting critiques of generated representations suggest a broader normative ambiguity: what *should* generative representations embody? In this work, we provide a conceptual framework that separates the descriptive aspects (what is instantiated by model outputs) from the normative (what we aspire for representations to depict and make possible). Drawing on longstanding work on modality in philosophy and linguistics, we characterize patterns of instantiation as implicit possibility assertions and use varieties of modality to structure distinct normative foundations for what representations should do. This framing reveals how the same representation may be evaluated differently depending on whether one prioritizes historical fidelity, conceptual possibility, aspirational aims, or other representational purposes. Rather than seeking a single “unbiased” representation, we argue for plurality in the normative expectations governing generative representations. We conclude by considering affordances for rendering these commitments explicit, critically engaging with them, deliberating over contested defaults, and providing users greater maneuverability over what is rendered possible.

Extended version —

<https://tinyurl.com/what-should-ai-generate>

Introduction

Generative AI (GAI) systems are increasingly celebrated for their ability to produce seemingly infinite, richly detailed outputs across domains such as entertainment, marketing, finance, scientific research, and healthcare (Ramdurai and Adhithya 2023). From a single prompt: “Give me a story about X” or “Generate an image of Y doing Z,” a model can produce a coherent narrative or detailed image without any further guidance.

This generative ease, however, masks numerous decisions required to produce such an output. Given a prompt “Give

me an image of a family,” the model must make numerous decisions: the number and configuration of people (nuclear, multi-generational, chosen family), their relationships and roles (parents, siblings, caregivers), and markers of economic status. Additional choices include racial/ethnic composition, ability/disability, housing conditions, and the inclusion or exclusion of non-nuclear members (grandparents, domestic workers, adult children). While some of these decisions remain ambiguous, others, such as which people are depicted, require a decision.

It seems uncontroversial that there exists no single perfect representation — one that, in totality, warrants no critiques and encompasses all possibilities. And yet, even when outputs from generative systems do not claim to be representative, there seems to be no shortage of critiques. In some cases, critiquing how generated outputs represent some groups in stereotypical and essentialist ways (Mack et al. 2024b; Qadri et al. 2023; Ghosh et al. 2024), while rendering others altogether invisible (Narayanan Venkit et al. 2023; Glazko et al. 2024a; Srinivasan and Uchino 2021). In other cases, heterogeneity has violated social expectations in troubling ways, spurring significant protest in depictions of Black and Asian Nazi soldiers (Vynck and Tiku 2024; Allyn 2024; Guo, Dong, and Hao 2024).

These critiques, though seemingly in tension, point toward distinct normative expectations. Should representations of historical contexts prioritize strict historical fidelity, or is historical plausibility sufficient? More broadly, how closely should generative outputs remain constrained by reality by default? Such choices may shape users’ ability to engage in speculative practices, including reasoning through counterfactuals and constructing counter-narratives. Attending to these choices, therefore, requires recognizing the plurality of purposes that representations can serve: they can depict not only what is, but also what is possible, permissible, or desirable, making alternative possibilities and broader moral lessons available for consideration. Disagreements over generated representations may thus arise not from a shared notion of “bias,” but from competing normative frameworks governing what representations should do. Whether a depiction is judged appropriate may depend on whether one prioritizes historical fidelity, conceptual possibility, aspirational equity, statistical typicality, or some other representational objective.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Generative systems make these consequential interpretive choices implicitly and arbitrarily, selecting normative frames without declaration or active steering. Unlike human communication, where these decisions can be clarified through dialogue (“What are you using this representation for? What do you want it to say?”), generative outputs resolve ambiguity silently, determining the relevant set of possibilities without users recognizing the contingency of what they encounter.

In this work, we contribute a unifying conceptual framework to separate the descriptive patterns rendered by generative systems from the variety of normative expectations regarding representations. By creating space for plurality among what models should depict, we delineate distinct evaluation axes for the same representation. We therefore shift from enumerating and mitigating biases to selecting the appropriate goals of representations. By asking “What purposes and possibilities do we want representations to embody?” we can separate the normative and descriptive notions of generative representations as free parameters and generated approaches to align them. Our proposed approaches span transparency mechanisms, deliberative structures, and user controls to offer a path toward enabling users to critically engage with instantiated patterns and align with their own normative goals.

Related Work

Much of the existing work on representational dynamics focuses on bias, or the (often unfair) favoring of a pattern of representations. This has been documented in GAI representations across many social groups such as gender bias (Kelly et al. 2025; Gupta et al. 2025b; Aneja, Gupta, and Vashistha 2025; Gupta et al. 2025a; Lin et al. 2025; Hada et al. 2024; Ovalle et al. 2025; Jourdan, Chevalier, and Favre 2025), disability bias (Glazko et al. 2024b; Mack et al. 2024a), political bias (Bang et al. 2024; Dai et al. 2025), ethnicity and sociodemographic biases (Li, Shirado, and Das 2025; Lin et al. 2025), linguistic biases (Gupta et al. 2025a; Lyu et al. 2025), religious biases (Seth et al. 2025), and mental health (Moore et al. 2025). In cases of markedness, patterns have been studied among a variety of granularities: from individual character caricatures (Cheng, Piccardi, and Yang 2023), to personas (Cheng, Durmus, and Jurafsky 2023; Mickel et al. 2025), to broader narratives and cultures (Gillespie 2024; Li et al. 2024).

Technical approaches to debiasing aim to remove or unlearn undesired patterns embodied in training data (Lin et al. 2024; Dixon et al. 2018; Gardner et al. 2021; Le Bras et al. 2020; Swayamdipta et al. 2020; Oren et al. 2020; Wu et al. 2021; Ross, Marasović, and Peters 2021; Asai and Hajishirzi 2020; Li et al. 2023; Kaneko and Bollegala 2021; Garg et al. 2019; Bolukbasi et al. 2016; Sedoc and Ungar 2019). Yet, continued debate remains regarding whether the effects of debiasing can be fully reversed (Gonen and Goldberg 2019; Prost, Thain, and Bolukbasi 2019), particularly in light of cases where debiasing might be undesirable, such as social science and medicine (McFadden et al. 1992; Obermeyer et al. 2019; Back et al. 2010). Work has critiqued broader practices of definitions for “good models” (Ganesh, Taik,

and Farnadi 2025), “misabstractions” (de Troya et al. 2025; Wang 2025), and narrowness in definitions of bias and accountability gaps (Ghosh and Wilson 2025; Hota and Jokinen 2025). Altogether, these characterizations of bias and debias are often defined relative to a default or dominant group. Comparative notions of fairness thus rely on a group to compare against, conflating parity for the full set of possibilities.

In response, recent years have yielded growing calls for “thickness” contextuality, situatedness, and richness both in the evaluators and resulting empirical evaluations (Qadri et al. 2025; Bennett, Kane, and Harrington 2025; Magomere et al. 2025; Mack et al. 2024b; Qadri, Mirowski, and Denton 2025; Khan, Casper, and Hadfield-Menell 2025). This has been composed of participatory approaches (Delgado et al. 2023; Birhane et al. 2022; Qi and Yu 2025; Dantec and DiSalvo 2013; Delgado, Barocas, and Levy 2022), value-sensitive design (Le Dantec, Poole, and Wyche 2009; Friedman 1996; Friedman, Kahn, and Borning 2002), and community-centered work (Bennett, Kane, and Harrington 2025; Bergman et al. 2024; Haimson et al. 2020; Li and Wu 2024; Wang et al. 2025; Li et al. 2025; Qadri et al. 2023). These works often target inclusion of a more diverse set of stakeholders, positioning particular groups’ lived experiences as key to informing key design and deployment decisions.

But representations have been shown to hold influence over a variety of spaces. Ranging from dynamics of relative visibility (Gillespie 2024), conceivability (Valkenburg 2024), reasoning (Fazelpour 2021; Chien and Danks 2024), future imagining (Oldenburg and Papyshev 2025), and oppressive normality (Lin and Huang 2025). Representations are the battlegrounds for their effects on conceptions and beliefs through phenomena such as symbolic annihilation (Gerbner 1970, 1972; Tuchman 2000), identity formation (Liu et al. 2017; Hogg and Hornsey 2007; Schmader, Croft, and Whitehead 2014), and normative beliefs, expectations, and behaviors (Bicchieri 2016; Paluck 2009; Chong and Ferrara 2009; Rogers et al. 1999). Generated representations should therefore be evaluated not only for what they depict, but also for how they shape broader dynamics of conceivability, recognition, and normality. Because these dynamics implicate multiple communities, values, and ways of understanding the world, no single community can necessarily serve as the sole arbiter of representational meaning. This motivates a plurality of normative expectations for generative representations—both regarding what should be possible to generate and what should be generated by default.

Recent work has made calls for greater “normative rigor” (Olteanu et al. 2025), with specific examinations of norms relative to specific groups or for particular purposes, such as according to queer and disability perspectives (Taylor et al. 2025; Bennett, Kane, and Harrington 2025), for moral norms (Sachdeva and van Nuenen 2025), legal applications (Hota and Jokinen 2025). This work does not attend to the potential for interaction or conflict between multiple normative foundations. We echo calls for greater attention to normative foundations, drawing on the study of modality to

structure engagement with broader sets of possibilities and move from a universal normative foundation to one that is actively selected or engaged with by a user.

Throughout this work, we draw on the longstanding study in philosophy, logic, and linguistics known as *modality*, where it has long been used to structure reasoning about possibility, likelihood, and necessity. In classical modal logic, statements are evaluated not solely for their truth or falsity but in relation to what could, must, and cannot be true under a given set of assumptions (Chagrov and Zakharyashev 1997; Chellas 1980; Garson 2023; Kripke 1963; Kripke et al. 1980). Modal logic provides a formalism for reasoning about possibilities: rules and conditions for gradations of possibility to explore the logical space of alternatives (Lewis 1973, 1986) (see **From Possible Worlds to Possibility Assertions**). In linguistics, modality plays a similarly foundational role in world construction. Natural language encodes modality through modal verbs (e.g., must, might, may, can), adverbs (e.g., necessarily, possibly), and syntactic structures that signal a speaker's stance toward a proposition (Portner 2009; Palmer 2001; Von Stechow 2005; Lyons 1977). Here, modal language not only conveys epistemic confidence or normative force but also structures how interlocutors interpret what is presented as inevitable, optional, desirable, or impossible. It grounds modal meaning in communicative context and shared background assumptions (Kratzer, Eikmeyer, and Rieser 1981; Kratzer 1991). In this work, we draw on varieties of modality, or distinct kinds of possibilities, to delineate normative frameworks and expectations for representations.

Given the interdisciplinary audience that this work aims to reach and feedback on the prevalent usage of “modality” to refer to modes of inputs and outputs in Computer Science, we henceforth minimize references to modality in favor of “possibility.” This means that terms that would otherwise be more precise for those familiar with the study of modality such as “modal stance,” “modal distance,” and “potential modal impression” (i.e., influence over modal beliefs) are often replaced with simpler terminology such as “influence/effect on possibilities.”

What is Depicted: Characterizing Instantiated Possibilities

We first provide structure to characterize what generative systems actually produce. Through patterns in instantiation and embodiment, generative systems transform concepts into implicit statements about what is possible, likely, and necessary.

Identifying Key Constructs and Their Patterns

First, we consider units of analysis broadly spanning entities, roles, and concepts specified by the prompt or embodied in the output. Entities are tangible, distinct objects and actors, roles are social or functional positions that actors occupy (e.g., nurse, teacher, firefighter), and concepts are abstract, relational, or contextual elements that inform the scene (e.g., a school, a classroom, educational practices,

healthcare systems, cultural norms, societal composition, social hierarchies).

Explicit, Implicit and Indeterminate Elements Derived from Dennett's (1989) terminology regarding mental representations, we propose characterizing attributes of elements according to their specificity: *explicit*, *implicit*, and *indeterminate* elements. This allows us to compare explicitness between the input and the output, and therefore distinguish what has been requested and imputed, and in what ways. The comparison surfaces the often-invisible or implicit decisions made in seamless generation, enabling evaluation, critical engagement, and intentional choice.

Explicit elements include those directly stated in the input or output. Implicit elements encompass characteristics that are not strictly required but rendered nonetheless. Implicit features facilitate analysis of broader cultural or historical subtext that may inform representational patterns.¹ Indeterminate elements involve ambiguous characteristics: those that are left unaddressed based solely on the output. Though seemingly trivial and not without its own challenges, the inclusion of this category aims to encompass characteristics that remain unspecified despite their potential cultural or contextual significance.

Assessing Variation Next, we examine the patterns in which these characteristics are selected. For a given prompt, a model's output may not be deterministic. Thus, the ways in which a characteristic remains constant or is highly variable inform understanding of implicit model patterns of possibility, particularly in resolving ambiguity.

To capture the comparative salience of these choices, we introduce the notion of *relative stability*: how consistently a characteristic appears across outputs, both explicitly and implicitly. Characteristics that remain constant across outputs exhibit *high stability*, while those that vary widely exhibit *high variability*. This axis of analysis helps identify which features the system treats as fixed versus flexible.

Altogether, we can conduct analysis across the stability axis in conjunction with specificity. The cross product of these dimensions (implicit, explicit) and (stable, variable) characterizes the representational patterns in system outputs. Comparing what is not requested with what is nonetheless embedded in outputs renders explicit the assumptions and interpretive choices introduced by instantiation. Explicit stable features that are explicitly in the prompt may simply reflect compliance with the user's input requests, thereby justifying the corresponding frequency. Stable features not explicitly requested in the prompt, however, reveal opportunities for alternative possibilities. These are the patterns that we aim to render legible (and eventually negotiable) for their potential normative underpinnings.

¹For simplicity and scope, we do not delve into the extent to which implied characteristics can be determined in inputs or outputs. We acknowledge that this is a sufficiently complex problem; however, there are many ways in which the set of implied characteristics or relations is likely not infinite and may be extracted from an appropriately scoped cultural context or history, such as through documented cultural attitudes and values.

Possibility Structure

From Possible Worlds to Possibility Assertions Modal logic distinguishes between different degrees of possibility through quantification over cognitive conceptions of all possible instantiations, i.e., “possible worlds,” where the world as we know it is just one of infinitely many possible worlds. Necessary features are intrinsic to the definition of a concept and essential to satisfy any instantiation of said concept. These are universal across the set of all possible worlds. Features that do not exist in all possible worlds can be subdivided into two categories: likely and possible, which we differentiate by their relative representation across all possible worlds. Likely features exist in many or most possible worlds, whereas possible features exist in one or more possible worlds (Menzel 2024; Phillips, Luguri, and Knobe 2015; Kment 2024; Von Fintel 2005).

We treat each instantiated output as a possible world, translating patterns in instantiation into possibility assertions (i.e., interpretive summaries that characterize the bounds of possibility rendered by system outputs). Characteristics that appear consistently across outputs (high stability) may be interpreted as necessary; those that appear frequently as likely or probable; those that appear occasionally as possible; and those that never appear as effectively impossible within the system’s representational space. These are not claims about what a model “believes” nor renders impossible to generate, but rather what they render visible through regularities in instantiation.

For example, consider a system that generates images of a nurse that consistently depict individuals in clinical hospital settings (high stability, explicit), wearing scrubs and attending to patients in beds (high stability, implicit), who are predominantly women, white or Asian, and young to middle-aged (high stability, implicit). We can articulate statements such as: “Nurses must work in hospitals” (due to consistent co-occurrence with clinical settings, excluding community health workers, school nurses, or home care providers), “Nurses are likely to be women” (due to statistical prevalence in outputs, marginalizing male nurses despite their growing presence in the profession), and “Nurses of other races, ages, or practice settings are possible but rare” (due to low frequency of, say, elderly Black nurses in community clinics or male nurses in pediatric care). These statements characterize the underlying structure of possibilities that the system outputs implicitly communicate about the space of possibilities.

These patterns may reinforce occupational gender stereotypes, narrow conceptions of healthcare work, and constrain what roles viewers perceive as accessible or typical for different demographic groups (Chien and Danks 2024). Such representational patterns echo the kind of gendered assumptions about caregiving work that the United States Supreme Court found unconstitutional in *Weinberger v. Wiesenfeld* (1975) (Asserson 1975). The Court struck down Social Security policies that presumed only women could be caregivers and recognized that such assumptions, even when statistically common at the time, functioned to reinforce and perpetuate gender-based occupational segregation rather than merely reflecting natural or inevitable divisions of la-

bor. In what follows, we argue that these possibility assertions are laden in all interactions with representations, underscoring the need for critical engagement and active deliberation over which kinds of possibility assertions are implicitly embedded in outputs.

What Should Be: A Human-Centered Approach for Effects of Representations

Having established a descriptive framework for characterizing what generative systems convey through embodied outputs, we now turn to the normative question: what *should* these systems depict? We provide conceptual means for reasoning about what should be represented by generated outputs. Beyond technical considerations of the diversity of these outputs, how might the ways that humans interact with (and are shaped by) representations inform the kinds of distributions and effects we desire to achieve through them?

Possibility Reasoning in Human Communication

For those unfamiliar with philosophical and linguistic uses of modality, first we argue that everyday questions and statements are already modally loaded. That is, they contain (implicit or explicit) assumptions about what the relevant set of possibilities are. While the appropriate or reasonable set of possibilities may go unstated when interlocutors share aligned assumptions, misalignment reveals both the consequences of implicit modal assumptions and what can be gained by rendering a model’s stance explicit.

Consider the following questions and their corresponding normative sets of possibilities:

1. I ask Bill Nye, the Science Guy, if the sky could be a different color. I expect him to respond based on physical possibilities given Earth’s atmospheric composition and the physics of light scattering, not metaphysical possibilities (if we collectively named these wavelengths differently) or biological counterfactuals (if humans lacked S-cone cells for perceiving blue wavelengths).
2. I ask my friend, “Where should we go to dinner tonight?” and expect answers constrained by practical feasibility: real restaurants within reasonable proximity to our location, open during our intended dining time. If they respond with locations only accessible via airplane or suggest fictional establishments, they violate the practical set of possibilities implicitly invoked.
3. I ask a kindergarten class, “What do you want to be when you grow up?” receiving answers like doctor, firefighter, astronaut, and president. I expect occupations constrained by cultural familiarity and social legibility, not metaphysical reinterpretations such as “happy” or “flexible.”
4. I ask a Doctor Who online forum whether The Doctor could regenerate into a non-human form. Responses navigate multiple nested normative frames simultaneously: what is canonically possible within the fictional universe (Time Lords can change appearance at will), what is practically likely given production constraints (unlikely

due to special effects costs and reduced actor expressiveness), and what is narratively desirable. The fandom engages in sophisticated reasoning over the appropriate set of possibilities across fictional, practical, and aesthetic dimensions at once.

In each scenario, both participants engage in implicit possibility reasoning: which possibilities are appropriate and relevant by default, relatively close (but perhaps stretching the boundaries), or warrant further or explicit specification. Whether I have to specify feasible constraints on restaurants or disclose options that may be at the normative bounds of these possibilities depends on how each party perceives the relevant set of possibilities and that of each other to be.

These examples reveal several critical aspects of how humans reason about the appropriate set. First, possibility reasoning is ubiquitous but not always explicit. When I ask Bill Nye about sky color, I do not preface my question with, “considering only physical and atmospheric constraints,” yet this framing is mutually understood. Second, successful communication requires alignment between participants’ assumptions about relevant possibilities. I do not specify to my friend that dinner suggestions should exclude restaurants requiring international travel, yet we share understanding of the operative constraints at play.

Third, and most importantly, humans use rich contextual cues to calibrate possibility expectations. We assess who is speaking (Bill Nye’s expertise suggests physical/scientific framing), why they are asking (dinner planning invokes practical feasibility), what domain we are in (kindergarten career aspirations invoke cultural norms about legitimate occupations), and what relationship we have (online fandom discussion permits playful exploration of possibility sets) to determine these factors. When someone violates expected implicit framing by suggesting a restaurant in another country, particularly without disclosing so, it feels wrong or even counterproductive even when technically accurate. The lack of disclosure may even implicitly signal misunderstanding in context or deliberate provocation.

Fourth and finally, possibility calibration is dynamic and negotiable. If my friend’s international restaurant recommendations are followed by, “I’m serious, I just got a bonus and want to celebrate,” I can recalibrate my understanding of the relevant set of possibilities. The key is that this calibration happens through contextual cues, interpersonal estimates, and appropriate disclosure.

Kinds of Possibility

The intuitions demonstrated above can be systematized through *varieties of modality*, or distinct categories grounded in different kinds of constraints (e.g., knowledge, norms, physical laws, or goals) rather than as a single all-encompassing set of possibilities (Kratzer, Eikmeyer, and Rieser 1981; Kratzer 1991). Kment (2024) posits that possibility and necessity are not single, universal categories, but rather, something that can be dependent on the kinds of constraints or rules one assumes. Modal claims (possibility statements) are therefore relational: rendered true or false relative to a specific basis for evaluation or “vari-

ety” (Copeland 2002; Kment 2024). Each variety provides important guidance on how we interpret and analyze the potential impact on our own conceptions of possibility for a given representation. In what follows, we provide a brief overview of a few varieties of modality.

The epistemic modality evaluates statements relative to what is known or believed. A proposition is epistemically possible if it is compatible with the available evidence and epistemically necessary if it must be true given that evidence. For example, “It might rain tomorrow” is epistemically possible if current weather data allows for rain, while “The sun must rise tomorrow” is epistemically necessary given current astronomical knowledge. This captures how possibility depends on information states or available evidence. If someone has never experienced rain at that time of the year, then the notion that “It might rain tomorrow” might be inconceivable, whereas having seen it rain perhaps the day before, the same person might consider it possible.

The metaphysical modality concerns what could or must be the case in virtue of the nature or identity of things. A proposition is metaphysically necessary if it cannot be otherwise without contradiction (e.g., “All squares must have four sides”), while it is metaphysically possible if it does not violate the essential properties or conceptual structure of the entities involved. This variety grounds reasoning about conceptual constraints and definitional properties.

Other varieties further broaden the landscape of sets of possibilities by considering normative constraints instead. For instance, the deontic modality evaluates statements relative to norms, laws, or moral rules, focusing primarily on permission and obligation. Under this modality, something is necessary if it ought to happen (e.g., “Drivers must stop at red lights”) and possible if it is permitted (Bybee and Fleischman 1995). The teleological modality evaluates necessity and possibility relative to goals or purposes: a feature is necessary if it is required to achieve an end and possible if it contributes to it.

We provide this brief introduction to varieties of modalities to highlight some of the distinctions between different kinds of possibilities. Unlike the previously provided examples where human intuitions of the relevant modalities are relatively clear, next we turn to cases of modal ambiguity. We separate differing critiques by their implicit modal variety, surfacing opportunities for critical evaluation and deliberation.

Which Possibility? Ambiguity in Generative Contexts

Beyond cases where the appropriate kinds of possibility seem uncontroversial or universal, many generative contexts present cases of genuine ambiguity – reasonable disagreement about which kinds of possibilities (i.e., which normative aims and standards) should govern what is generated. The associated disagreement stems not from technical misinterpretation but from undefined expectations about what and whose values, purposes, and conceptions of possibility *should* shape generated representational decisions and defaults.

Consider a case where GAI is used to produce a script for the next episode of a popular medical television show. Upon instantiation, a model must determine how realistic and accurate medical procedures and scenarios must be: are they limited to current best medical practices or do they involve future speculative miracle fixes? Considering now the kinds of possibilities centers the broad spectrum available: a medical educator might argue that we should never depict anything other than best practices (epistemic modality), as this might cause additional harms by non-medical professionals in emergency scenarios or impede the learning of future medical students. A TV writer or even a medical researcher, however, might take a more open-ended approach: citing that the show aims to inspire, enable escapism, and motivate new research alternatives (teleological modality). The same prompt thus yields not only different outputs but outputs premised on fundamentally different normative baselines for permissible effects of the representation. This distinction may be elided or even uncritically accepted without the framework provided in this work. Moreover, a bias evaluation may not capture this disagreement: as both outputs could appear equally unbiased (particularly relative to social representations) while encoding fundamentally different assumptions about which possibilities ought to be represented.

What Varieties Clarify Rather than treating conflicting critiques as undifferentiated claims of bias, modal varieties reveal that disputants may be operating from distinct normative baselines. Consider the contentious case in which a user prompts Gemini for images of “the Pope,” and receives depictions of a Black man and a woman in papal vestments (see **Figure 1**). Under the epistemic modality, all documented popes have been men, asserting that the Pope must be male and most likely white.² According to the deontic modality grounded in current Church law, however, a Black male pope is permissible, as race constitutes no formal constraint; indeed, Black cardinals have been considered potential candidates in the two most recent papal conclaves.

Depictions of female popes, in contrast, violate ecclesiastical law prohibiting female ordination. This means that according to the deontic and epistemic modalities, such depictions are not permissible. However, the metaphysical modality in which the pope is defined as the spiritual leader of the Catholic Church renders both Black and female popes possible. From a teleological modality with a goal of challenging gender exclusion, female papal representations may even be necessary for demonstrating institutional reform.

The disagreement between a devout Catholic and a feminist critic about these depictions is therefore not simply a dispute around historical accuracy but a debate over which kinds of possibility should govern by default. A devout Catholic might view female papal depictions not merely as inaccurate but as sacrilegious, treating Church doctrine as authoritative in determining the boundaries of what should be represented as possible (De Boer 2016; Waldron 2012). A feminist critic, however, might argue that depicting a female pope challenges exclusionary norms and enables cri-

²While three potentially African popes existed, their race remains unverified (Browne 1998).



Figure 1: Image of a pope attributed to @EndWokeness on X and Frank J. Fleming, produced by Gemini on 02/21/2024. (Left) A dark-skinned middle-aged woman with features suggesting East or Southeast Asian ancestry. (Right) A dark-skinned middle-aged man appearing to be of African descent.

tique of patriarchal institutional practices (Okin 2005). What one community experiences as legitimate critique, another experiences as offensive and inconceivable. Naming these varieties does not resolve the disagreement but reframes it – elucidating opportunities for structured deliberation over which normative baseline is appropriate for which representational purposes. For reasons of space, see additional **Case Studies 1 & 2** in the appendix of the extended version).

Defensible Groundings This surfaces new questions: how might we select and defend particular kinds of possibilities as default? How might we navigate different normative perspectives when interacting with the same output? Normative arguments are never purely deductive, so formal logic alone cannot determine which kinds should govern. As a start, defensibility can be assessed by the degree to which a kind of possibility coheres with widely shared normative frameworks such as universal human rights (Assembly 1949), constitutional freedoms (Rozinski 2025), and democratic values (Thomassen 2007). Drawing on these existing frameworks carries cross-contextual legitimacy precisely because they do not presuppose any single community’s particular commitments but collective ones. A normative baseline defensible under these frameworks constitutes a reasonable initial standpoint to adopt because it engages a shared normative vocabulary rather than encoding one group’s standards as universal.

For the papal example, one might argue that since depictions of a female pope do not violate an individual Catholic’s freedom of religion (which protects their right to hold and practice their beliefs, not to have those beliefs reflected in generative defaults), such representations can be defensibly rendered possible. A Catholic user remains free to specify “male pope” or “historically accurate pope” to specify modes aligned with Church doctrine, but the default need not encode theological hierarchies as universal constraints. Given broader commitments to gender equality and non-discrimination, Black and female papal renderings are thus defensibly possible under unspecified prompting.

These broad anchors connect naturally to domain-specific principles that lend further support to particular normative groundings. Aspirational defaults under the deontic modality find grounding in anti-subordination principles and capability expansion (Colker 2017; Balkin and Siegel 2002; Nussbaum 2011; Sen 2001), counter-narrative scholarship (Delgado et al. 2023; Lundholt, Maagaard, and Piekut 2018; Grabe and Dutt 2015; Bruneau, Lane, and Saleem 2017; Potvin et al. 2023; Solórzano and Yosso 2002; Kinloch, Penn, and Burkhard 2020), and epistemic justice frameworks (Fricker 2007). Historically-constrained defaults under the epistemic modality draw support from documentary fidelity and principles of historical accuracy that serve functions of preserving evidence and avoiding revisionism (Streich 2002; Temin and Dahl 2017).

These competing foundations explain why reasonable disagreement persists: multiple sets of possibilities can each claim legitimate grounding in different normative traditions. The structure outlined in this work does not adjudicate between them conclusively, but instead renders explicit which foundational commitments and corresponding conceptions of harm are doing the normative work. How, for example, should one weigh harms arising from patriarchal exclusion against perceived harms to religious freedom or the integrity of sacred institutions? Our framework does not resolve such questions, but clarifies where the disagreement lies. Ultimate decisions therefore require negotiation between competing values, balancing critical awareness of these decisions against fundamental steerability.

Reconciling the Descriptive and the Normative: Ways to Bridge the Gap

Having separated the descriptive patterns and normative foundations, our framework reveals several key tensions: how to address misalignment between what models encode and what users expect or desire; how to navigate disagreements among users, viewers, and diverse stakeholder groups with different normative expectations; and how to design appropriate controls and agency in navigating various notions of possibility. The strategies below provide a prescriptive approach, treating both descriptive patterns and normative targets as free parameters that determine the appropriate affordances. Rather than offering a single “right” answer, we highlight the contours of these challenges as we consider four interrelated approaches.

Descriptive Disclosure: Is Transparency Enough?

One approach to addressing normative misalignment centers on transparency, aiming to inform users about what patterns generative models codify and affording opportunities for critical engagement. Extant examples include a message beneath the prompt box reading, “ChatGPT can make mistakes. Check important info.” to more detailed approaches, such as proposed by Huang et al., providing distributions of output characteristics (such as the gender breakdown of generated doctors) alongside comparison concepts (such as how those distributions shift between prompts for “a doctor” versus “a nurse”).

While this represents meaningful progress towards informing users, it lacks connection between descriptive patterns and normative implications. A user made aware that “92% of generated nurses are depicted as women” does not tell them what to do with this information: whether this reflects reality, historical precedent, media convention, or something else entirely.

We propose a transparency mechanism that makes this normative dimension explicit by contrasting AI output distributions with multiple modal baselines simultaneously. As illustrated in **Figure 2**, for a given prompt such as “Generate some images of a nurse,” a transparency panel might display how key characteristics such as gender, age, and affect are distributed across AI outputs alongside their corresponding rates in Reality (epistemic), Historical Depictions, and Contemporary Media. The “Notable Divergences” panel summarizes key points where an AI output significantly diverges or aligns with any baseline, directing user attention toward potential modal influence or risk of departure from the user’s own normative expectations. For instance, while gender representation in AI outputs (92% female) bears close resemblance to occupational reality (88%), age representation diverges sharply: AI outputs depict nurses as young 78% of the time compared to 45% in reality, reflecting media conventions more than reality. Making these comparisons visible does not prescribe what the user should conclude, but provides the conceptual scaffolding to ask: which baseline is relevant to my purpose and how do the AI’s default modal stance compare?

This strategy assumes that if users understand a model’s descriptive tendencies relative to distinct or even their own normative frames, they can better calibrate their expectations and interpretations of outputs accordingly. Transparency does not interfere with existing generative workflows, only supplementing interactional paradigms. By shifting the burden onto users to evaluate their own goals and aims against those disclosed to be embedded by the system, transparency targets informed choice.

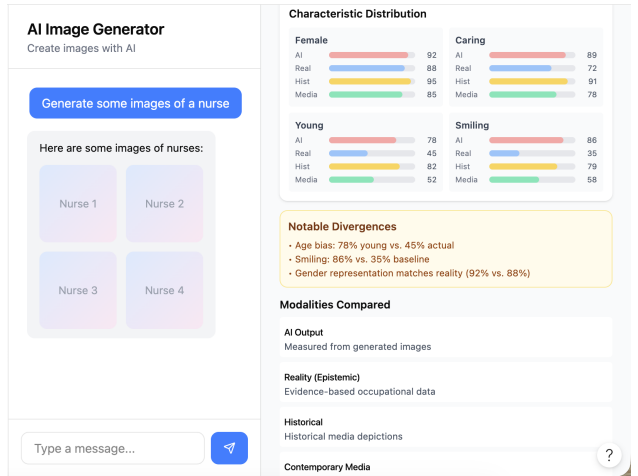


Figure 2: Proposed User Interaction Pane with Transparency Mechanisms Alongside

However, disclosure presumes users have the capacity and context to meaningfully engage with information about model behavior (Bo et al. 2024; Brody et al. 2024; Metzger et al. 2024). Even with explicit normative baselines, a user confronted with the statistic that “78% of generated nurses are depicted as young versus 45% in reality” must still determine what to do with that information: which baseline is relevant to their purpose, is the deviation problematic for their intended use case, and what action, if any, is warranted (e.g., additional prompting or generations). For users with limited statistical literacy or domain knowledge, the disconnect between what is conveyed and what can be meaningfully operationalized may be substantial. Additionally, surfacing all divergences at every interaction risks disclosure fatigue. A more targeted approach might notify users only when AI outputs diverge notably from baselines of social significance, with user values rather than system defaults determining salience (Sutton and Wood 2025). For this, future work might explore more accessible representations, such as visual features or even output order that guide users toward the representations most salient to their stated goals.

Furthermore, transparency does not address the irreversibility of representational exposure: once outputs are viewed, read, or heard, they cannot be simply unseen. A user who sees a stereotypical generated image cannot undo that exposure by reading an accompanying disclaimer. Disclosure may allow an opportunity for critical engagement in coarse ways, such as reducing engagement with the platform itself, but cannot ensure such boundaries are respected once representations are taken out of their generating contexts (e.g., on social media or broadcast news). A user particularly impacted by a generated depiction may not find that accompanying information attenuates cognitive effects, suggesting the need for something preempting exposure (Favela and Amon 2023; Hancock et al. 2020). In these cases, disclosure still leaves broader publics vulnerable to resulting representational harms (i.e., symbolic annihilation) and provides limited opportunities for individuals to influence what should be depicted. Transparency does not address cases in which users fundamentally disagree with what the model treats as descriptively accurate. Disclosure that a model reproduces historical patterns of occupational gender segregation, for instance, does little to address whether such reproduction is acceptable; it merely makes the problem explicit. Users seeking normatively desirable outputs rather than descriptively accurate ones gain little from knowing that the system encodes patterns they find objectionable. Actionable changes to their requests may be intuitive in cases with clear counterparts, but may be ambiguous in others, leaving unclear mechanisms for recourse for users.

Reconstructing GAI around Human Intuitions for What Should Be

An alternative approach involves restructuring generative systems to instead better align with human normative intuitions for possibilities. A combination of technical and design strategies can operationalize this approach by making normative choices more explicit and negotiable in the generation process.

A few quick questions to give you more tailored guidance:

Want to be notified when Claude responds? Notify

What's your current financial situation? < 1 of 3 >

- 1 Just starting out (little savings)
- 2 Some savings, no investments
- 3 Already investing, want to expand
- 4 Significant assets to manage

Something else Skip

Or reply directly...

Sonnet 4.6

Figure 3: Claude prompts for additional information for ambiguous prompt, “Help me plan my finances.”

Leveraging our proposed framework, platforms could provide users with explicit “modes” of generation that encode distinct normative frameworks. A “historical mode” might privilege epistemic accuracy and documentary fidelity, generating outputs that reflect actual historical distributions and constraints. A “training data mode” could generate variation purely based on proportional representation in the training corpus, making the model’s learned patterns explicit. A “media mode” might reflect patterns represented in mainstream media. A “creative mode” might emphasize outputs that intentionally deviate from reality, ranging across unverified possibilities and fantastical imaginations. These modes could be further specified into sub-categories drawing from existing genres (science fiction, historical fiction, documentary), aesthetic styles (realism, surrealism, abstraction), intentional metaphysical subversions (counterfactuals, alternative histories), or other salient axes of variation relevant to particular domains. Importantly, interaction modes can dictate distributional constraints on outputs rather than dictating individual instantiations, allowing for variation within coherent normative frameworks. Requiring that users actively select a mode can spark critical engagement and acknowledgment of the different kinds of possibilities and normative foundations available, providing a compromise between cognitive burden (to specify the relevant normative commitments) and system-imposed decisions.

Existing approaches in this vein do not directly target distinctions in kinds of possibility. Upon account creation, Claude prompts users to input what kind of work they do, selecting across roles such as legal, student, educator, consultant, or researcher, ostensibly for tailoring starter prompts and system behavior. Before answering certain open-ended questions, Claude provides options for users to specify additional context or constraints (see Figure 3). While asking for additional information provides one way to increase understanding of what a user intends, follow-up questions may still operate within a single underlying normative foundation, asking a user to clarify their financial goals, for example, does not surface whether they want outputs grounded in best practice, accounting for potential financial hardships, or

other major life decisions (e.g., having a child or throwing a big wedding). Follow-up questions such as these may implicitly center certain normative foundations (i.e., epistemic) while leaving others unexplored.

Other alternatives employ personas, which similarly aim to specify relevant values that influence outputs based on human-legible roles, professions, and social and demographic categories. Users might prompt “as a historian” or “as a diversity advocate” to signal which considerations should be prioritized. While intuitive for users unfamiliar with various conceptions of possibility, it remains unclear whether invoking a persona implicitly constrains outputs to the deontic modality or requires additional decisions about an individual acting in that role. Khan, Casper, and Hadfield-Menell (2025) observed a high level of instability, incoherence between evaluated versus held-out cultural dimensions, and erratic behavior under prompt steering.

A combination of persona-based and variety-specific constraints may provide user-accessible modes of interaction with consistent possibility constraints. A “judge” persona might operate primarily within the deontic modality (concerned with what is legally established), while a “surrealist artist” persona operates within a metaphysically relatively unconstrained modality (exploring what is conceptually possible regardless of empirical likelihood). A “policy advocate” might invoke deontic modality (what should be according to the law), while a “data scientist” invokes epistemic modality (what is statistically probable given observed data). Crucially, making these normative commitments explicit and legible elucidates the open challenge of ensuring users meaningfully understand these limitations, rather than drawing on what users theorize “a historian” or “a diversity advocate” prioritize, value, or enact in practice.

Consensus Making vs Personalization: Opportunities for Deliberation Creating modes or mode-persona profiles surfaces tensions around determining the appropriate normative baselines. Does “what should be depicted” have to be agreed upon at a group consensus level, community-specific level, or even an individual level?

Consensus-oriented approaches seek to establish shared defaults – outputs that align with widely negotiated normative commitments and serve as common starting points across diverse user populations. Consensus appeals in its potential to create universal interactional environments where all users encounter the same default behaviors, fostering shared representational commons and mutual intelligibility in how systems operate. However, the papal example illustrates significant challenges: even a seemingly straightforward question (e.g., what the Pope looks like) surfaces irreconcilable normative frameworks across communities, suggesting that normative agreement may be impossible in some domains. For consequential social issues: forcing a consensus risks oversimplifying the default set of possibilities, potentially constraining users to engage only within predetermined normative frameworks rather than the full range of normatively defensible positions.

Democratic and participatory approaches such as citizen juries, stakeholder assemblies, or structured public

consultation offer mechanisms for navigating these challenges (Smith and Wales 2000; Sandover, Moseley, and Devine-Wright 2021; Muller and Kuhn 1993; Schuler and Namioka 1993; Farias, Bendor, and Van Eekelen 2022). Such processes aim to balance the attentional and critical engagement demands of high-stakes applications with genuine plurality of perspectives, creating opportunities for contested values to be surfaced and negotiated. Yet even well-designed participatory processes have their trade-offs, requiring substantial resources and challenges of scale and specificity across the full diversity of use contexts (Ghoshal and Dasgupta 2023; Miller et al. 2007).

Personalization offers an alternative: systems that adapt to individual or community-specific normative frameworks, generating value-aligned outputs for different users even under identical prompts. A Catholic school using a generative system for educational materials might receive outputs that default to historically accurate papal depictions, while a feminist media organization using the same system for a project on institutional reform might receive outputs that foreground aspirational representations. This approach affords experiential plurality in the use of generative representations, allowing users to interact with systems that respect their particular normative frameworks. Community-specific versions might encode locally appropriate normative commitments, enabling different cultural contexts to shape system behavior according to their values.

However, personalization, particularly when optimized for seamless normative inference that anticipates user preferences without explicit specification, risks downstream fragmentation (Magin et al. 2022). If different populations receive systematically different generated realities, will this exacerbate epistemic polarization, produce filter bubbles, and fracture shared understanding? When users inhabit personalized information environments that confirm rather than challenge their existing normative commitments, opportunities for encountering alternative perspectives may diminish.

Deliberative approaches may offer a middle path through concrete institutional mechanisms. Tiered governance structures could establish general principles through broad-based citizen assemblies, domain-specific defaults through affected community participation, and individual adjustments through user preference settings. Standing advisory boards comprising diverse stakeholders could review and update mode defaults periodically, responding to emerging concerns and shifting social contexts (Newman et al. 2011). Public comment periods modeled on regulatory rule-making could solicit feedback on proposed default changes, with companies required to publish response documents explaining how input shaped final decisions (Morrell 2013). Contestation mechanisms could allow users or communities to flag specific behaviors as problematic, triggering review processes with transparent decision criteria and public rationales (Gorwa, Binns, and Katzenbach 2020). Possibility influence impact assessments conducted before deployment could evaluate how different default choices affect various groups, similar to privacy impact assessments (Wright 2012). Experimental deployments could test competing defaults with different user cohorts, gathering empirical evi-

dence about user satisfaction, viewer comprehension, critical empowerment, and downstream effects before committing to system-wide defaults.

These mechanisms acknowledge that deliberation is an ongoing process of negotiation as contexts, technologies, and social norms evolve. Open challenges remain in designing processes that are genuinely inclusive (not dominated by a vocal majority), responsive (able to adapt to rapid technological change), and practically feasible. Success likely requires experimentation with multiple deliberative forms, transparency about their limitations, and willingness to revise approaches as we learn which mechanisms effectively accommodate disagreement without demanding false consensus.

Non-Normative Use: Fundamental Rights to Controllability

Finally, modes offer just one possibility along the axis of controllability and maneuverability. Given the broad set of applications for which artists, researchers, educators, and everyone in between may wish to use generative AI, more fine-grained controls may be required. Existing paradigms pose prompting as one lever of control. However, it is unclear that prompting necessarily nor always guarantees meaningful forms of controllability (Lam et al. 2026; Zamfirescu-Pereira et al. 2023).

Mack et al. (2024b), Bennett, Kane, and Harrington (2025), and Qadri, Mirowski, and Denton (2025) have demonstrated cases in which outputs do not satisfy the direct requests of users, despite iterative processes and alternative terminology. For example, users attempting to generate images of disabled scientists or gender-nonconforming executives may find that models systematically resist or distort such combinations, reverting to more statistically familiar configurations regardless of how the prompt is refined. These limitations raise questions about fundamental rights to representation: when models prove incapable of producing particular combinations or forms of representations, such as certain body types, cultural practices, counterstereotypical pairings, are users entitled to “generability” as a matter of principle?

The concept of generability builds on notions of representational justice. If certain subjects, identities, or configurations remain systematically difficult or impossible to generate, this technical limitation may indirectly entrench or exacerbate existing patterns of visibility and erasure. When a model cannot readily produce images of disabled scientists, Indigenous mathematicians, or gender-nonconforming executives, it has the potential to perpetrate “quality-of-service harms” (Chehbouni et al. 2024).

This raises additional questions around fairness in ease of generation: if a system can generate “a male nurse,” should it equally be capable of generating “a female construction worker,” with the failure to do so constituting a form of unfairness? This raises an equally important normative question surrounding limitations (or lack thereof) on generability. Should systems be capable of generating any requested representation, including those that might be used to create non-consensual imagery, facilitate fraud, misinforma-

tion, or propagate harmful stereotypes? The tension between enabling legitimate representational diversity and preventing malicious use complicates straightforward appeals to generability rights.

Further complicating these questions is the interaction between generability constraints and generation modes. When operating under “historical mode,” should a model refuse to generate a female Roman general, with the same refusal under “creative mode” constituting a quality of service harm? Or should the system generate such images but accompany them with warnings about historical implausibility? The framework set forth in this work reveals that generability constraints cannot be universal, that what constitutes an appropriate limitation depends on the normative framework governing generation. A request that violates the commitments of one mode (historical accuracy, demographic proportionality, stereotype avoidance) might be perfectly legitimate under another. Determining which normative framework should govern in a given context, however, is a question of governance, requiring collective deliberation over the values and trade-offs that generative systems should embody. We do not prescribe universal constraints on generability, but explain why any such constraints must be grounded in explicit normative commitments.

Conclusion

In this work, we deconstruct the space of bias and critiques of generative representations into two parts: the descriptive (what is embodied by model outputs) and the normative (what we aspire for from representations). We introduce varieties of modalities to broaden and structure the dimensionality of evaluations around generated representations. We discuss key considerations around resulting affordances, spanning transparency, default modes and deliberation, and fundamental rights to controllability. These are motivated by ideas that we must first render explicit the underlying patterns, which we can then critically engage with, deliberate over, and design clear maneuverability mechanisms (either through individual control, collective governance, or standards for utility). All of these hinge on the parsing of normative framings of the aims of representations, asking first, “What should be embodied by generated representations?”

Limitations

In this work, we only discuss a few of the possible varieties of modality. Varieties themselves are not fixed sets of possibilities, and even those listed above may not be easily defined by a fixed set of rules. This does not mean that they lack utility nor undermine the contributions of our work. Each variety surfaces a set of normative claims about what is possible, likely, and necessary, which in turn creates opportunities to evaluate, communicate, and negotiate their relevance. In addition, we do not provide immediately operationalizable solutions in the prescriptive section; we acknowledge that the development of such practical tools is an important research direction, likely requiring extensive research in human-computer interaction, user-interface design, collective governance, and public engagement and education initiatives.

Acknowledgments

Thank you to Sina Fazelpour, Maria De-Arteaga, David Danks, Justin Shin, Arjun Subramonian, Jovy Chan, Alexander Mussnug, Emma Duncan, Ben Mylius, Seliem El-Sayed, Linda Eggert, Leif Wenar, Anne Newman for their feedback throughout the process and without which this work would not be what it is today.

References

- Allyn, B. 2024. Google races to find a solution after Ai Generator Gemini misses the mark.
- Aneja, U.; Gupta, A.; and Vashistha, A. 2025. Beyond Semantics: Examining Gender Bias in LLMs Deployed within Low-resource Contexts in India. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 2784–2795. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Asai, A.; and Hajishirzi, H. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5642–5650.
- Assembly, U. N. G. 1949. *Universal declaration of human rights*, volume 3381. Department of State, United States of America.
- Asserson, B. M. 1975. Weinberger v. Wiesenfeld. *Hofstra L. Rev.*, 4: 149.
- Back, S. E.; Payne, R. L.; Simpson, A. N.; and Brady, K. T. 2010. Gender and prescription opioids: Findings from the National Survey on Drug Use and Health. *Addictive behaviors*, 35(11): 1001–1007.
- Balkin, J. M.; and Siegel, R. B. 2002. The American civil rights tradition: Anticlassification or antisubordination. *Issues Legal Scholarship*, 2: i.
- Bang, Y.; Chen, D.; Lee, N.; and Fung, P. 2024. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11142–11159.
- Bennett, C. L.; Kane, S. K.; and Harrington, C. N. 2025. Toward Community-Led Evaluations of Text-to-Image AI Representations of Disability, Health, and Accessibility. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 256–270.
- Bergman, S.; Marchal, N.; Mellor, J.; Mohamed, S.; Gabriel, I.; and Isaac, W. 2024. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports*, 14(1): 6616.
- Bicchieri, C. 2016. *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.
- Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.
- Bo, J. Y.; Kumar, H.; Liut, M.; and Anderson, A. 2024. Disclosures & disclaimers: investigating the impact of transparency disclosures and reliability disclaimers on learner-LLM interactions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 12, 23–32.
- Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.
- Brody, G.; Rouillard, V.; Asherov, D.; and Aravind, A. 2024. Evaluating the Effectiveness of AI Source Disclosure in Human–AI Communication.
- Browne, M. 1998. The three african popes. *The Western Journal of Black Studies*, 22(1): 57.
- Bruneau, E.; Lane, D.; and Saleem, M. 2017. Giving the underdog a leg up: a counternarrative of nonviolent resistance improves sustained third-party support of a disempowered group. *Social psychological and personality science*, 8(7): 746–757.
- Bybee, J.; and Fleischman, S. 1995. Modality in grammar and discourse: An introductory essay. *Modality in grammar and discourse*, 14: 503–517.
- Chagrov, A.; and Zakharyashev, M. 1997. *Modal logic*. Oxford University Press.
- Chehbouni, K.; Roshan, M.; Ma, E.; Wei, F.; Taik, A.; Cheung, J. C. K.; and Farnadi, G. 2024. From representational harms to quality-of-service harms: A case study on llama 2 safety safeguards. In *Findings of the Association for Computational Linguistics: ACL 2024*, 15694–15710.
- Chellas, B. F. 1980. *Modal logic: an introduction*. Cambridge university press.
- Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. *arXiv preprint arXiv:2305.18189*.
- Cheng, M.; Piccardi, T.; and Yang, D. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. *arXiv preprint arXiv:2310.11501*.
- Chien, J.; and Danks, D. 2024. Beyond behaviorist representational harms: A plan for measurement and mitigation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 933–946.
- Chong, A.; and Ferrara, E. L. 2009. Television and divorce: Evidence from Brazilian novelas. *Journal of the European Economic association*, 7(2-3): 458–468.
- Colker, R. 2017. Anti-subordination above all: A disability perspective. In *Disability and Equality Law*, 177–206. Routledge.
- Copeland, B. J. 2002. The genesis of possible worlds semantics. *Journal of Philosophical logic*, 31(2): 99–137.
- Dai, S.; Cao, Z.; Wang, W.; Pang, L.; Xu, J.; Ng, S. K.; and Chua, T.-S. 2025. Media source matters more than content: Unveiling political bias in llm-generated citations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 17267–17287.

- Dantec, C. A. L.; and DiSalvo, C. 2013. Infrastructuring and the formation of publics in participatory design. *Social Studies of Science*, 43(2): 241–264.
- De Boer, D. 2016. Picking up the Pieces: Catholic Material Culture and Iconoclasm in the Low Countries.
- de Troya, I. n.; Kernahan, J.; Doorn, N.; Dignum, V.; and Dobbe, R. 2025. Misabstraction in Sociotechnical Systems. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 1829–1842. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Delgado, F.; Barocas, S.; and Levy, K. 2022. An uncommon task: Participatory design in legal AI. *Proceedings of the ACM on human-computer interaction*, 6(CSCW1): 1–23.
- Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23.
- Dennett, D. C. 1989. *The intentional stance*. MIT press.
- Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.
- Farias, P. G.; Bendor, R.; and Van Eekelen, B. F. 2022. Social dreaming together: A critical exploration of participatory speculative design. In *Proceedings of the Participatory Design Conference 2022-Volume 2*, 147–154.
- Favela, L. H.; and Amon, M. J. 2023. Enhancing Bayesian approaches in the cognitive and neural sciences via complex dynamical systems theory. *Dynamics*, 3(1): 115–136.
- Fazelpour, S. 2021. Norms in counterfactual selection. *Philosophy and Phenomenological Research*, 103(1): 114–139.
- Fricker, M. 2007. Epistemic injustice: Power and the ethics of knowing.
- Friedman, B. 1996. Value-sensitive design. *interactions*, 3(6): 16–23.
- Friedman, B.; Kahn, P.; and Borning, A. 2002. Value sensitive design: Theory and methods. *University of Washington technical report*, 2(8): 1–8.
- Ganesh, P.; Taik, A.; and Farnadi, G. 2025. Systemizing multiplicity: The curious case of arbitrariness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 1032–1048.
- Gardner, M.; Merrill, W.; Dodge, J.; Peters, M. E.; Ross, A.; Singh, S.; and Smith, N. A. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1801–1813.
- Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 219–226.
- Garson, J. 2023. Modal Logic. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition.
- Gerbner, G. 1970. Cultural indicators: The case of violence in television drama. *The Annals of the American Academy of Political and Social Science*, 388(1): 69–81.
- Gerbner, G. 1972. Symbolic functions. *Television and social behavior: Media content and control*, 1: 28.
- Ghosh, S.; Venkit, P. N.; Gautam, S.; Wilson, S.; and Caliskan, A. 2024. Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach. *arXiv preprint arXiv:2407.14779*.
- Ghosh, S.; and Wilson, K. 2025. Bias Is a Math Problem, AI Bias Is a Technical Problem: 10-Year Literature Review of AI/LLM Bias Research Reveals Narrow [Gender-Centric] Conceptions of ‘Bias’, and Academia-Industry Gap. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 1091–1106.
- Ghoshal, S.; and Dasgupta, S. 2023. Design Values in Action: Toward a Theory of Value Dilution. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2347–2361.
- Gillespie, T. 2024. Generative AI and the politics of visibility. *Big Data & Society*, 11(2): 20539517241252131.
- Glazko, K.; Mohammed, Y.; Kosa, B.; Potluri, V.; and Mankoff, J. 2024a. Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 687–700.
- Glazko, K.; Mohammed, Y.; Kosa, B.; Potluri, V.; and Mankoff, J. 2024b. Identifying and Improving Disability Bias in GPT-Based Resume Screening. FAccT '24, 687–700. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Gonen, H.; and Goldberg, Y. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614.
- Gorwa, R.; Binns, R.; and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945.
- Grabe, S.; and Dutt, A. 2015. Counter narratives, the psychology of liberation, and the evolution of a women’s social movement in Nicaragua. *Peace and Conflict: Journal of Peace Psychology*, 21(1): 89.
- Guo, X.; Dong, L.; and Hao, D. 2024. RETRACTED: cellular functions of spermatogonial stem cells in relation to JAK/STAT signaling pathway. *Frontiers in Cell and Developmental Biology*, 11: 1339390.
- Gupta, I.; Joshi, I.; Dey, A.; and Parikh, T. 2025a. “Since Lawyers are Males..”: Examining Implicit Gender Bias in Hindi Language Generation by LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 3254–3264. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.

- Gupta, I.; Joshi, I.; Dey, A.; and Parikh, T. 2025b. “Since Lawyers are Males.”: Examining Implicit Gender Bias in Hindi Language Generation by LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25, 3254–3264. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Hada, R.; Husain, S.; Gumma, V.; Diddee, H.; Yadavalli, A.; Seth, A.; Kulkarni, N.; Gadiraju, U.; Vashistha, A.; Seshadri, V.; and Bali, K. 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*, FAccT ’24, 1926–1939. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Haimson, O. L.; Gorrell, D.; Starks, D. L.; and Weinger, Z. 2020. Designing trans technology: Defining challenges and envisioning community-centered solutions. In *proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13.
- Hancock, P.; Kaplan, A.; MacArthur, K.; and Szalma, J. L. 2020. How effective are warnings? A meta-analysis. *Safety Science*, 130: 104876.
- Hogg, M. A.; and Hornsey, M. J. 2007. Self-concept threat and multiple categorization within groups. In *Multiple social categorization*, 128–152. Psychology Press.
- Hota, A.; and Jokinen, J. P. 2025. NomicLaw: Emergent Trust and Strategic Argumentation in LLMs During Collaborative Law-Making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 1278–1289.
- Huang, Y.; Hanwen Deng, W.; Xiao, S.; Eslami, M.; Hong, J. I.; Narechania, A.; and Perer, A. 2026. Vipera: Blending Visual and LLM-Driven Guidance for Systematic Auditing of Text-to-Image Generative AI. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–18.
- Jourdan, F.; Chevalier, Y.; and Favre, C. 2025. FairTranslate: an English-French Dataset for Gender Bias Evaluation in Machine Translation by Overcoming Gender Binarity. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25, 150–166. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Kaneko, M.; and Bollegala, D. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1256–1266.
- Kelly, M.; Tahaei, M.; Smyth, P.; and Wilcox, L. 2025. Understanding Gender Bias in AI-Generated Product Descriptions. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25, 2587–2615. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Khan, A.; Casper, S.; and Hadfield-Menell, D. 2025. Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25, 2151–2165. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Kinloch, V.; Penn, C.; and Burkhard, T. 2020. Black lives matter: Storying, identities, and counternarratives. *Journal of Literary Research*, 52(4): 382–405.
- Kment, B. 2024. Varieties of Modality. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition.
- Kratzer, A. 1991. Modality. In von Stechow, A.; and Wunderlich, D., eds., *Handbuch Semantik*, 639–50. De Gruyter.
- Kratzer, A.; Eikmeyer, H.-J.; and Rieser, H. 1981. The notional category of modality. *Formal semantics: The essential readings*, 289–323.
- Kripke, S. A. 1963. Semantical considerations on modal logic. *Acta philosophica fennica*, 16.
- Kripke, S. A.; et al. 1980. *Naming and necessity*, volume 217. Springer.
- Lam, M. S.; Shaikh, O.; Xu, H.; Guo, A.; Yang, D.; Heer, J.; Landay, J. A.; and Bernstein, M. S. 2026. Just-In-Time Objectives: A General Approach for Specialized AI Interactions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–26.
- Le Bras, R.; Swayamdipta, S.; Bhagavatula, C.; Zellers, R.; Peters, M.; Sabharwal, A.; and Choi, Y. 2020. Adversarial filters of dataset biases. In *International conference on machine learning*, 1078–1088. Pmlr.
- Le Dantec, C. A.; Poole, E. S.; and Wyche, S. P. 2009. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1141–1150.
- Lewis, D. 1973. Counterfactuals and comparative possibility. In *IFS: Conditionals, belief, decision, chance and time*, 57–85. Springer.
- Lewis, D. 1986. *On the plurality of worlds*, volume 322. Oxford Blackwell.
- Li, H.; Jiang, L.; Dziri, N.; Ren, X.; and Choi, Y. 2024. CULTURE-GEN: Revealing Global Cultural Perception in Language Models through Natural Language Prompting. *arXiv preprint arXiv:2404.10199*.
- Li, J.; Li, Q.; Gong, R.; Wang, L.; and Wu, S. 2025. Our Collective Voices: The Social and Technical Values of a Grass-roots Chinese Stuttered Speech Dataset. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2768–2783.
- Li, Q.; and Wu, S. 2024. ”I Want to Publicize My Stutter”: Community-led Collection and Curation of Chinese Stuttered Speech Data. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2): 1–27.
- Li, Y.; Du, M.; Wang, X.; and Wang, Y. 2023. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14254–14267.

- Li, Y.; Shirado, H.; and Das, S. 2025. Actions Speak Louder than Words: Agent Decisions Reveal Implicit Biases in Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 3303–3325. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Lin, T.-a.; and Huang, L. T.-L. 2025. AI, Normality, and Oppressive Things. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 1574–1574.
- Lin, Z.; Guan, S.; Zhang, W.; Zhang, H.; Li, Y.; and Zhang, H. 2024. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9).
- Lin, Z.; Trogrlic, G.; de Vreese, C.; and Helberger, N. 2025. Dangerous Criminals and Beautiful Prostitutes? Investigating Harmful Representations in Dutch Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 1005–1014. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Liu, P.; Zhao, F.; Zhang, B.; and Dang, Q. 2017. Small change makes a big splash: The role of working self-concept in the effects of stereotype threat on memory. *The Journal of Psychology*, 151(7): 613–631.
- Lundholt, M. W.; Maagaard, C. A.; and Piekut, A. 2018. Counternarratives. *The international encyclopedia of strategic communication*, 1–11.
- Lyons, J. 1977. *Semantics*, volume 2. Cambridge university press.
- Lyu, H.; Luo, J.; Kang, J.; and Koenecke, A. 2025. Characterizing Bias: Benchmarking Large Language Models in Simplified versus Traditional Chinese. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 2815–2846. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Mack, K. A.; Qadri, R.; Denton, R.; Kane, S. K.; and Bennett, C. L. 2024a. “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*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.
- Mack, K. A.; Qadri, R.; Denton, R.; Kane, S. K.; and Bennett, C. L. 2024b. “They only care to show us the wheelchair”: disability representation in text-to-image AI models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–23.
- Magin, M.; GEIB, S.; Stark, B.; and Jürgens, P. 2022. Common core in danger? Personalized information and the fragmentation of the public agenda. *The International Journal of Press/Politics*, 27(4): 887–909.
- Magomere, J.; Ishida, S.; Afonja, T.; Salama, A.; Kochin, D.; Foutse, Y.; Hamzaoui, I.; Sefala, R.; Alaagib, A.; Dalal, S.; Marchegiani, B.; Semenova, E.; Crais, L.; and Hall, S. M. 2025. The World Wide recipe: A community-centred framework for fine-grained data collection and regional bias operationalisation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 246–282. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- McFadden, A. C.; Marsh, G. E.; Price, B. J.; and Hwang, Y. 1992. A study of race and gender bias in the punishment of school children. *Education and treatment of children*, 140–146.
- Menzel, C. 2024. Possible Worlds. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition.
- Metzger, L.; Miller, L.; Baumann, M.; and Kraus, J. 2024. Empowering calibrated (dis-) trust in conversational agents: A user study on the persuasive power of limitation disclaimers vs. authoritative style. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Mickel, J.; De-Arteaga, M.; Liu, L.; and Tian, K. 2025. More of the Same: Persistent Representational Harms Under Increased Representation.
- Miller, J. K.; Friedman, B.; Jancke, G.; and Gill, B. 2007. Value tensions in design: the value sensitive design, development, and appropriation of a corporation’s groupware system. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work*, 281–290.
- Moore, J.; Grabb, D.; Agnew, W.; Klyman, K.; Chancellor, S.; Ong, D. C.; and Haber, N. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 599–627. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Morrell, E. 2013. Public comment periods and federal environmental impact statements: Potentials and pitfalls from the American experience. *Michigan Journal of Sustainability*, 1.
- Muller, M. J.; and Kuhn, S. 1993. Participatory design. *Communications of the ACM*, 36(6): 24–28.
- Narayanan Venkit, P.; Gautam, S.; Panchanadikar, R.; Huang, T.-H.; and Wilson, S. 2023. Unmasking nationality bias: A study of human perception of nationalities in ai-generated articles. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 554–565.
- Newman, S. D.; Andrews, J. O.; Magwood, G. S.; Jenkins, C.; Cox, M. J.; and Williamson, D. C. 2011. Community advisory boards in community-based participatory research: a synthesis of best processes. *Preventing chronic disease*, 8(3): A70.
- Nussbaum, M. C. 2011. Creating Capabilities: The human development approach. Cambridge, MA: Belknap Press of Harvard.
- Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

- Okin, S. M. 2005. 'Forty acres and a mule' for women: Rawls and feminism. *Politics, Philosophy & Economics*, 4(2): 233–248.
- Oldenburg, N.; and Papishev, G. 2025. The Stories We Govern By: AI, Risk, and the Power of Imaginaries. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 1939–1950.
- Olteanu, A.; Blodgett, S. L.; Balayn, A.; Wang, A.; Diaz, F.; Calmon, F. d. P.; Mitchell, M.; Ekstrand, M.; Binns, R.; and Barocas, S. 2025. Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor. *arXiv preprint arXiv:2506.14652*.
- Oren, I.; Herzig, J.; Gupta, N.; Gardner, M.; and Berant, J. 2020. Improving compositional generalization in semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2482–2495.
- Ovalle, A.; Pavasovic, K. L.; Martin, L.; Zettlemoyer, L.; Smith, E. M.; Chang, K.-W.; Williams, A.; and Sagun, L. 2025. The Root Shapes the Fruit: On the Persistence of Gender-Exclusive Harms in Aligned Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25*, 3094–3105. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Palmer, F. R. 2001. *Mood and modality*. Cambridge university press.
- Paluck, E. L. 2009. Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *Journal of personality and social psychology*, 96(3): 574.
- Phillips, J.; Luguri, J. B.; and Knobe, J. 2015. Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, 145: 30–42.
- Portner, P. 2009. *Modality*. OUP Oxford.
- Potvin, G.; Hazari, Z.; Khatri, R.; Cheng, H.; Head, T. B.; Lock, R. M.; Kornahrens, A. F.; Woodle, K. S.; Vieyra, R. E.; Cunningham, B. A.; et al. 2023. Examining the effect of counternarratives about physics on women's physics career intentions. *Physical Review Physics Education Research*, 19(1): 010126.
- Prost, F.; Thain, N.; and Bolukbasi, T. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 69–75.
- Qadri, R.; Diaz, M.; Wang, D.; and Madaio, M. 2025. The case for "thick evaluations" of cultural representation in ai. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 2067–2080.
- Qadri, R.; Mirowski, P.; and Denton, R. 2025. AI and Non-Western Art Worlds: Reimagining Critical AI Futures through Artistic Inquiry and Situated Dialogue. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–17.
- Qadri, R.; Shelby, R.; Bennett, C. L.; and Denton, E. 2023. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 506–517.
- Qi, X.; and Yu, J. 2025. Participatory design in human-computer interaction: Cases, characteristics, and lessons. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–26.
- Ramdurai, B.; and Adhithya, P. 2023. The impact, advancements and applications of generative AI. *International Journal of Computer Science and Engineering*, 10(6): 1–8.
- Rogers, E. M.; Vaughan, P. W.; Swalehe, R. M.; Rao, N.; Svenkerud, P.; and Sood, S. 1999. Effects of an entertainment-education radio soap opera on family planning behavior in Tanzania. *Studies in family planning*, 30(3): 193–211.
- Ross, A.; Marasović, A.; and Peters, M. E. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3840–3852.
- Rozinski, T. 2025. Constitutional Freedoms in the United States.
- Sachdeva, P.; and van Nuenen, T. 2025. Normative Evaluation of Large Language Models with Everyday Moral Dilemmas. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25*, 690–709. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Sandover, R.; Moseley, A.; and Devine-Wright, P. 2021. Contrasting views of citizens' assemblies: Stakeholder perceptions of public deliberation on climate change. *Politics and Governance*, 9(2): 76–86.
- Schmader, T.; Croft, A.; and Whitehead, J. 2014. Why can't I just be myself? A social cognitive analysis of the working self-concept under stereotype threat. *Social Psychological and Personality Science*, 5(1): 4–11.
- Schuler, D.; and Namioka, A. 1993. *Participatory design: Principles and practices*. CRC press.
- Sedoc, J.; and Ungar, L. 2019. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the first workshop on gender bias in natural language processing*, 55–61.
- Sen, A. 2001. Economic development and capability expansion in historical perspective. *Pacific Economic Review*, 6(2): 179–191.
- Seth, A.; Choudhury, M.; Sitaram, S.; Toyama, K.; Vashistha, A.; and Bali, K. 2025. How deep is representational bias in llms? the cases of caste and religion. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 2319–2330.
- Smith, G.; and Wales, C. 2000. Citizens' juries and deliberative democracy. *Political studies*, 48(1): 51–65.
- Solórzano, D. G.; and Yosso, T. J. 2002. Critical race methodology: Counter-storytelling as an analytical framework for education research. *Qualitative inquiry*, 8(1): 23–44.
- Srinivasan, R.; and Uchino, K. 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.

- Streich, G. W. 2002. Is there a right to forget? Historical injustices, race, memory, and identity. *New Political Science*, 24(4): 525–542.
- Sutton, J.; and Wood, M. M. 2025. Opting Out: Over-Alerting and Warning Fatigue in the Era of Wireless Emergency Alerts. *Journal of Contingencies and Crisis Management*, 33(3): e70076.
- Swayamdipta, S.; Schwartz, R.; Lourie, N.; Wang, Y.; Hajishirzi, H.; Smith, N. A.; and Choi, Y. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9275–9293.
- Taylor, J.; Mire, J.; Spektor, F.; DeVrio, A.; Sap, M.; Zhu, H.; and Fox, S. E. 2025. Un-Straightening Generative AI: How Queer Artists Surface and Challenge Model Normativity. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25*, 951–963. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Temin, D. M.; and Dahl, A. 2017. Narrating historical injustice: Political responsibility and the politics of memory. *Political Research Quarterly*, 70(4): 905–917.
- Thomassen, J. 2007. Democratic Values. In Dalton, R. J.; and Klingemann, H., eds., *The Oxford Handbook of Political Behavior*. Oxford University Press. ISBN 9780199270125.
- Tuchman, G. 2000. The symbolic annihilation of women by the mass media. In *Culture and politics: A reader*, 150–174. Springer.
- Valkenburg, G. 2024. An Enquiry Into Modes of Non-Existence. *Science, Technology, & Human Values*, 01622439241300744.
- Von Fintel, K. 2005. Modality and language.
- Vynck, G. D.; and Tiku, N. 2024. Why google took down Gemini’s AI image generator and the drama around IT - The Washington Post.
- Waldron, J. 2012. *The harm in hate speech*. Harvard University Press.
- Wang, A. 2025. Identities are not Interchangeable: The Problem of Overgeneralization in Fair Machine Learning. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25*, 485–497. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Wang, D.; Denton, R.; Sinha, A. K.; Sheth, S.; Wilcox, L.; Mustafa, M.; Eslami, M.; Holstein, K.; Sap, M.; Smith, A. D.; et al. 2025. Unanticipated Lessons from Communities: Navigating Society-Centered Research in the AI Era. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–3.
- Wright, D. 2012. The state of the art in privacy impact assessment. *Computer law & security review*, 28(1): 54–61.
- Wu, T.; Ribeiro, M. T.; Heer, J.; and Weld, D. S. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6707–6723.
- Zamfirescu-Pereira, J. D.; Wong, R. Y.; Hartmann, B.; and Yang, Q. 2023. Why Johnny can’t prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–21.