Identifying Harm in Personalized, Generative AI Systems Requires User-Centered Auditing at the Interaction Level

Hannah Cha
Stanford University
Stanford, CA, USA
Microsoft Research
New York, New York, USA

Abstract

Personalized, generative AI systems increasingly adapt their behavior to individual users over time, fundamentally changing model behavior. While existing auditing approaches have been effective at surfacing harms in non-personalized contexts, they often rely on static, simulated evaluations and definitions of harm that aggregate across broad, group categories. In this position paper, we argue that such approaches can fail to capture emergent harms in personalized generative AI systems, where harms emerge through interpretations of ongoing interaction and evolve with user history. We identify three presuppositions underlying many harm auditing paradigms: that harms can be (1) specified outside real-world interaction, (2) defined non-pluralistically within groups, and (3) treated as static. One might argue that personalized systems could simply learn definitions of what constitutes as harm to individual users through repeated interactions. However, we argue that attempts to surface user harms through deeper personalization risk imposing asymmetric burdens of labor and privacy on marginalized users. Consequently, we propose reframing understandings of harm as adaptive, user- and community-centered processes, and outline design directions that shift auditing from retrospective evaluation toward infrastructures that support ongoing articulation of harm in interaction. Our work highlights the need for auditing and design practices that better reflect the pluralistic and evolving nature of harm understanding in personalized generative AI systems.

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods; HCI theory, concepts and models**.

Keywords

end-user auditing, harms from biases, user-centered design

ACM Reference Format:
Hannah Cha. 2026. Identifying Harm in Personalized, Generative AI Systems Requires User-Centered Auditing at the Interaction Level. In *Proceedings of Human-Centered Evaluation and Auditing of Language Models Workshop at CHI 2026 (HEAL @ CHI '26).* ACM, New York, NY, USA, 9 pages.

HEAL @ CHI '26, Barcelona, Spain
2026.

1 Introduction

Generative AI systems are being increasingly integrated into everyday applications, such as chatbots [20, 62] and writing assistants [32, 60]. These systems have increasingly grown to personalize, or adapt outputs to individual users [23], based on past behavior [65, 99] and preferences [27, 40]. These methods promise more relevant and engaging interactions for users [33, 41]. While personalization is not new in computational systems [38, 47, 61], the nature of personalization in generative AI is fundamentally different. Unlike selection-based personalization, which chooses from a fixed set of outputs, generative AI reshapes the distribution of possible outputs, producing indeterminate, context-dependent, and evolving behaviors [42, 65]. In fact, personalization can fundamentally alter the nature of model behavior itself [92]. As personalization continues to shape the ways that systems interact with users, it is crucial to be able to document exactly when and how harms may emerge.

Existing paradigms to audit harms in generative AI systems often occur as non-personalized, simulated evaluations [35, 74, 75], where the model does not retain memory of any prior interactions and the simulated context can be abstracted away from real world use. While these approaches have systematically revealed biases in baseline models [35, 44, 55, 74], they may fail to capture true user harms arising as a result of personalized interaction. Harm cannot only be determined by what the output constitutes, but from the way individual users experience and interpret these outputs [10, 58, 90]. As AI systems become personalized to individual users, merely auditing for harms in non-personalized contexts risks missing emergent harms that impact real users. Additionally, current systems that users interact with provide few mechanisms for users to articulate harm, contest system assumptions, or directly shape how personalization unfolds over time. In the absence of such mechanisms, personalized AI systems default to making inferences about users [68]. Ultimately, this risks treating user preferences, such as what constitutes as harm, as something that can be inferred from proxies [92] rather than directly articulated by users.

In this position paper, we focus on *experiential harms*, which consist of harms that users can self-identify and report experiencing negative affect from [90]. We argue that personalized generative AI systems create experiential harms at the *interaction level* that are often difficult to surface through traditional auditing mechanisms. More specifically, we argue that many, general-purpose auditing approaches presuppose that:

(1) **Harms can be specified outside of real-world interaction**: Auditing often happens in a memory-less, non-personalized

contexts, abstracted away from how systems may truly harm users in the real world.

(2) **Harms can be defined non-pluralistically within groups**: Auditing often defines harm at broad, demographic categories (e.g. race, gender), which can obscure the pluralistic, intersectional ways members of the same group can experience harm.

(3) **Harms are static**: Many auditing methods lack continual or evolving evaluation over time, failing to capture how personalization introduce new harmful behavior over time, or how user definitions of harm may evolve.

We further argue that attempting to personalize definitions of harm to individual users with current system paradigms can create asymmetries of *labor* and *privacy*, disproportionately burdening users in minoritized groups. To address these challenges and bridge the gap in understanding true user harms, we propose design directions that acknowledge defining harm as an *adaptive, user- and community-centered process*, emphasizing evolving, pluralistic understandings of harm without causing disproportionate burden on users in minority groups. Ultimately, this work motivates the need for better harm auditing mechanisms that allows all users, especially those from minoritized backgrounds, to reap the benefits of personalized, generative technologies.

2 Related Work

2.1 Personalized, Generative AI Systems

Personalization is not new in computational systems. Recommender systems [47], search engines [38], and adaptive interfaces [61] have long tailored outputs to users. In these traditional systems, personalization primarily affects selection: which item is recommended [72], which result is ranked higher [38], or which predefined label is assigned [4]. Although users may interpret these outputs differently, the system's behavior itself often remains stable and enumerable. Thus, algorithmic decision rules can be probed to surface discrimination or harms because outputs are repeatable and enumerable across test inputs [75].

By contrast, generative AI systems produce outputs through stochastic generation that may be conditioned on user behavior and preferences [65, 99], prior conversational history [42], or even inferred attributes [74]. Thus, personalization in generative settings does not merely select among existing outputs, but reshapes the distribution over possible behaviors. The system may adopt different tones, frames, assumptions, and normative stances depending on the user it is interacting with [91]. In fact, personalization fundamentally changes model behavior, even modifying behavior on standard benchmarks [92]. Consequently, the model does not exhibit a single, stable behavior, but rather a family of behaviors that continually evolves in a personalization context that requires new ways of evaluation [37].

Recent work has attempted to better understand and evaluate how personalized interaction can affect behaviors for generative AI systems, such as chatbots [28, 79, 91]. We extend this need for new evaluation paradigms for personalized, generative AI systems specifically for the harms that users may experience.

2.2 Surfacing Harms Via Algorithm Auditing

Algorithm auditing describes the process of surfacing harms by directly analyzing the outputs of such systems [5, 43, 52]. Such harms include stereotypes [30], problematic output [5, 43], or surfacing disparities in outcome based on demographic group in domains such as healthcare [55, 66, 77], employment [13, 94, 97], and housing [3, 39, 106]. Many of these audit-based methods, especially in the context of generative AI systems, often consist of researchers simulating a use case of an LLM through prompts and iteratively evaluating a series of AI outputs. Potential harms from these systems are often surfaced by providing proxies for demographic groups, such as name [74] or dialect [35], and measuring discrepancies across groups. Such simulated studies have measured these discrepancies in the natural language used to describe groups [14, 35, 46], outcomes groups achieve [9, 74], or the way marginalized groups are portrayed in generative imagery [93, 100].

However, one common critique of these audit-based methods is that these evaluations are created and run by either AI practitioners [52, 75, 83] or AI researchers [11, 17, 24, 25, 70], who may be disconnected from the communities actually experiencing harm [7]. As a result, there has been a rise in end users and communities identifying emergent harms from AI systems [22, 78]. Users can discover nuanced and community-specific harms that have previously been overlooked by researchers or practitioners [50, 54, 69, 102, 104]. For instance, Deng et al. [21]'s WeAudit created an infrastructure allowing for users to audit amongst community members and report harms to AI practitioners in actionable ways. These user-centered auditing practices have been helpful in surfacing community-specific harms, such as stereotypical representations of non-Western cultures [30], inappropriate outputs in cultural contexts [69] and the lack of disability [50] and gender [29] representation in generative image models. Overall, these auditing methods focus on experiential harms, or harms that can be surfaced through self-reported experiences of negative affect [90], from the users themselves.

Nonetheless, existing auditing approaches can miss harms emerging from the ways that users interact with personalized AI tools. Simulated, stateless audits happen outside of true user interaction, and many community-based methods may consult communities at the design phase rather than continuing to evaluate the impact of tools after they are deployed [34]. Furthermore, even end-user auditing often functions with the goal of providing technical recommendations and remediations for AI practitioners [21, 56, 70, 73], rather than preventing harm or reshaping outputs for users as they occur. This can lead to an asymmetry of burden and benefit, where those who the work required by the system may not be the one who receive its benefits [31]. Ultimately, our work highlights this gap in auditing paradigms by outlining the nature of harm in personalized, generative AI systems that current auditing methods may fail to capture.

3 Why Existing Auditing Approaches Are Insufficient for Personalized Generative AI Systems

In this section, we argue that personalized, generative AI systems create harms that current auditing approaches may be insufficient

for surfacing. In this section, we outline how existing auditing approaches often presuppose harm (1) can be specified outside of real world interaction, (2) is non-pluralistic within groups, and (3) is static.

3.1 Auditing Approaches Presuppose Harm Can be Specified Outside Real-World Interaction

We argue that existing auditing approaches for defining and quantifying harm presuppose harms are specifiable outside interaction. Many auditing approaches have operationalized potential harms arising from AI systems through simulated studies. These simulated evaluations have quantified harms such as stereotyping [14, 35, 46, 51], unequal treatment [9, 74], or erasure of marginalized groups [93, 100]. While these strategies have enabled important advances in understanding harms from generative AI systems, they share an assumption that harms can be articulated and evaluated independently of how systems are experienced in regular use.

Personalized generative AI systems do not produce fixed or memoryless outputs. Rather, they generate responses dynamically through interaction, adapting to user behavior and preferences [65, 99] or prior conversational history [42]. As a result, harms cannot be fully characterized by isolated outputs or one-shot evaluations. Prior work has shown that such evaluations frequently fail to capture how systems truly behave in personalized contexts [92]. Consequently, outputs from simulated audits may not correspond to the harms users actually experience, while significant real-world harms may remain entirely invisible to evaluators. Recent real-world incidents illustrate how severe harms can emerge through prolonged, situated interaction rather than isolated prompts. While extreme, real world cases such as the generation of CSAM [63] or suicide [12] demonstrate consequential failure modes that may not be detectable through standard, static audit paradigms.

Thus, traditional evaluation of generative AI systems, often occurring in simulated, memoryless, or post-hoc audits, are abstracted from real-world use cases of these systems and its subsequent impact on users. Taken together, this suggests that existing auditing paradigms may fail to capture harms in personalized generative AI systems. By defining and evaluating harm outside of real-world interaction, these frameworks risk overlooking harms that emerge as systems adapt to users.

3.2 Auditing Approaches Presuppose Non-Pluralistic Definitions of Harm

Existing auditing approaches can further obscure asymmetries in how harm is distributed and experienced. Many fairness evaluations operationalize harm at an aggregate group level, commonly defined by broad demographic categories such as race or gender [71, 74, 84, 103]. However, these groupings to evaluate for harm implicitly assume that members of a group experience harms in similar ways. This assumption can mask misaligned, alienating, or harmful interactions experienced by subsets of users within a group. Even participatory approaches, aimed at directly understanding harms from groups themselves, can run the risk of obscuring asymmetries and complexities in how harm is experienced. Critical scholarship has long cautioned against equating participation with representation [34, 80]. Processes that elevate a small number of what may seem like representative voices, as in participatory design, may obscure internal disagreement, marginalize less powerful members, or reproduce dominant norms that exist within a community [16, 67, 101]. Prior work has highlighted the challenges of pluralistic alignment, where generative AI systems overlook human social diversity between groups [2, 81]; we extend this idea to within-group diversity and highlight the pitfalls of documenting harms and preferences at the group level. Importantly, we do not focus on non-negotiable harms, such as the generation of slurs, explicit stereotypes, or overtly abusive content, for which there is broad, normative agreement of what constitutes harm. Instead, we examine outputs that occupy an ambiguous middle ground, where framing, assumptions, or normative stance can be interpreted in fundamentally different ways depending on users' lived experiences.

Figure 1 illustrates how an example of an output from a personalized, generative AI system, while not overtly stereotypical or detectable as harmful in traditional auditing paradigms, can be interpreted in fundamentally different ways by users sharing a demographic identity. Two users within the same demographic group may interpret the response in two contrasting ways, each being rooted in their own lived experience. Psychological studies have documented this within-group pluralism of what constitutes as harm or bias [10], such as in the context of cultural appropriation [58]. Furthermore, as illustrated in the figure, additional identities (e.g. low-income status) can further shape how harm is perceived, experienced, or intensified. Thus, intersectionality adds another dimension of consideration for the multiplicity of interpretation of what can be determined as harmful: different forms of inequality manifest from various axes of identity (e.g. race, gender, class), creating unique experiences of privilege or discrimination [18]. Thus, trying to understand or audit for harms at broad group categories risk compounding harm by obscuring how multiple dimensions of identity and experience shape interpretation and definitions of harm [15, 16, 19, 67, 90]. Prior work has shown that, due to the combinatorial explosion of subgroups, measuring intersectional fairness exhaustively can be intractable [57]. Similarly, enumerating all plausible interpretations of harm for a given system output may be impossible; even the type of harm (e.g. racism, sexism, ageism) should neither be measured or mitigated in the same way [89].

The pluralistic ways users within a group can experience harm is a problem for current personalization mechanisms in generative AI systems, which often relies on identity-based proxies in attempts to personalize to users. For instance, Wang et al. [91] finds that GPT was more likely to recommend *Black Panther* to names associated with Black individuals and *Little Women* to names associated with women in personalized contexts, flattening the diversity of preferences of a group into stereotypes. Even if personalized systems were able to avoid explicit stereotypes, the fact that personalization can depend on identity-based proxies showcases how these systems can inadvertently elevate the preferences of a subset of users into implicit group standards, erasing the pluralistic ways of how harm can occur. Allowing for dominant members of groups to determine what constitutes as harm can lead to its own harms. Psychological studies have documented cases of lateral violence, where members of oppressed groups can marginalize members within the same group [95, 96]. Similarly, qualitative studies have

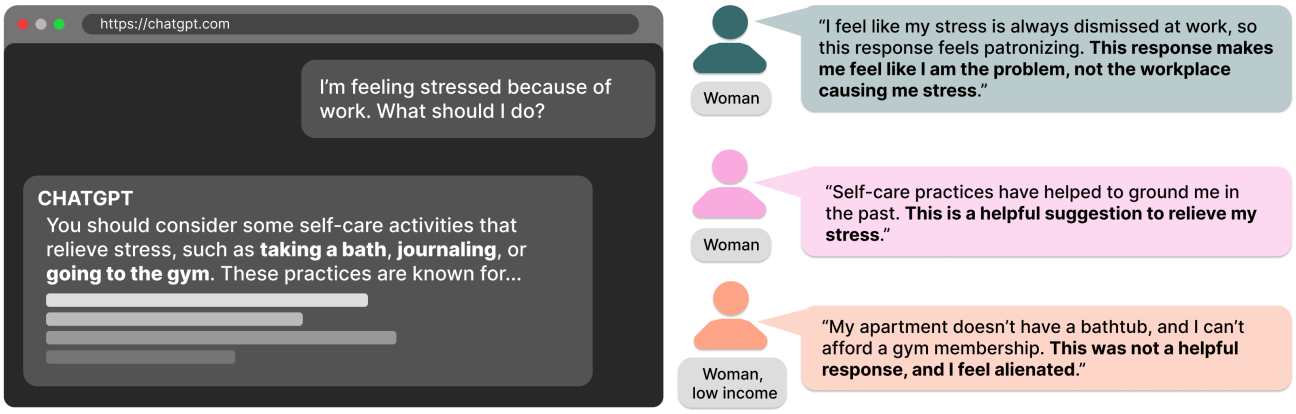

Figure 1: We present an example of case where plausible interpretation of a single chat output by two users in the same demographic group can cause harm to one and not the other. We also consider how intersectionality can affect the harm a user might experience, where additional identities (e.g. income status) can add new harm dimensions for a user in the same demographic group.

documented intragroup discrimination, where intersectional identities can result in harm within the same marginalized group [49]. Consequently, personalized generative systems operate under conditions of structural pluralism, where users who share demographic identities may nonetheless have legitimate, conflicting expectations of what constitutes harmful system behavior.

3.3 Auditing Approaches Presuppose Harm as Static

Furthermore, we argue that many existing auditing approaches fail to account for how harms can evolve over the course of interaction. Interpretations of what constitutes as harm may not be static even for a single individual. Psychology studies have found, for instance, that situational and dispositional factors can prime individuals to consider previously ambiguous statements as harmful [6] or as a microaggression [36]. Even an individual's level of historical knowledge affects their ability to perceive harms, such as racism [8]. Thus, it may be plausible, for example, that a user who initially appreciates culturally specific food recommendations may later come to experience them as reductive as they learn to recognize microaggressions, such as how food recommendations can be a mechanism for stereotyping. Thus, system behavior may shift from being perceived as helpful to harmful as users continue to evolve.

Traditional end-user auditing mechanisms rarely account for the interpretive fluidity of harm, treating user preferences and definitions of harm as stable rather than contextual and longitudinal. Empirical audits of large language models typically rely on one-off evaluations, such as static benchmark datasets [7], prompt sets [74], or simulated personas [76], to assess harmful behaviors. Thus, such audits may fail to capture how harms vary across time, interaction history, and changing user understanding. Even participatory approaches can fall into the trap of treating community-driven insights as static artifacts, often consulted at design time and translated into system requirements without long term engagement [34, 85]. Unless evaluation and community connection continues beyond the creation of the system itself, harms and insights from the community when they put the system to use can be missed [26, 34, 64].

Thus, in the context of personalized, generative AI, many auditing approaches and even participatory methods may fail to consider how harm might evolve over time as users themselves change or build history with personalization. Ultimately, without longitudinal audits or regular consultation of users or communities, it becomes difficult to account for how harms evolve as systems adapt, users change, and contexts shift.

4 Perfect Personalization Doesn't Close the Auditing Gap

A natural response to the limitations outlined above might be that if current personalization mechanisms don't adequately capture individual harm, then the solution could be deeper, more accurate personalization. One might argue that personalized, generative AI systems could eventually learn what constitutes as harm for a given user by personalizing based on inferred user discomfort or past negative signals. In theory, if users interact with personalized AI systems long enough, it may be able to perfectly capture fluid preferences and avoid harmful output. However, we argue that attempting to resolve this issue further personalization with the current paradigm of personalized generative AI systems leads to an asymmetry of **labor** and **privacy** for users in the minority.

4.1 Asymmetry of Labor

Creating a perfectly personalized, generative AI system for a given user begs the question of what labor it would take to create such an experience. Without explicit mechanisms for users to articulate harm or transparency into whether such articulation shapes future

behavior, users may be repeatedly exposed to harmful interactions before a system adapts to them. Prior work has already found that LLMs often default to outputs that cater towards WEIRD (Western, Educated, Industrialized, Rich, Democratic) perspectives and values [1, 76]. Thus, users who fall outside of that default would be more likely to encounter harmful outputs that need to be corrected. This dynamic distributes the labor of correction unevenly: users whose interpretations align with dominant norms may experience less friction for creating personalized experiences, while those whose experiences diverge must repeatedly intervene to correct system assumptions.

Additionally, even if users were to articulate harms as they occur, the lack of transparency in current AI systems [98] and the difficulty for users to steer AI system behavior [87] make it unclear whether this labor would prevent harms in future outputs. Often, the goal of AI auditing is technical remediation by AI practitioners or model developers [56, 70, 73]. However, this consolidates power in the hands of platforms, who can decide whether to address the harms users experience and how they should do so. Thus, users who document harms they experience directly to the AI system or through traditional auditing mechanisms are not guaranteed to experience less harm as a result of their labor.

Ultimately, repeated harmful outputs from the system can lead to reduced trust and adoption among marginalized groups, who have already been documented as less likely to be AI adopters [105]. This can ultimately exacerbate digital divides [88] in who benefits from personalized AI systems, concentrating the advantages of adaptive systems among users whose interpretations align more closely with dominant norms. Thus, minoritized users may encounter more harms and bear disproportionate labor in making their preferences legible to the system, creating a labor asymmetry.

4.2 Asymmetry of Privacy

In addition to imposing unequal labor burdens, attempts to achieve perfect personalization in generative AI systems may also create an asymmetry of privacy. Personalized, generative AI systems often rely on the collection and retention of user data to adapt system behavior [45, 48]. However, the amount and sensitivity of information required to avoid harms is not evenly distributed across users. For users whose preferences and interpretations of harm already align with dominant norms embedded in system defaults, effective personalization may require little additional disclosure. Their needs may be more likely to be met by generic system behavior, allowing them to benefit from the system without explicitly revealing sensitive aspects of their identity, experiences, or values. In contrast, users whose preferences diverge from these defaults may have to disclose more information to correct system assumptions. To steer system behavior away from alienating or harmful outputs, they may have to reveal information that is more personal, sensitive, or identity-linked than what is required of users of dominant groups. Even without explicitly revealing aspects of their identity, personalized systems may infer their identity based on their discomfort or disagreement with system defaults.

Many marginalized communities, especially Indigenous communities, are already concerned about how data collection and disclosures can become a form of extraction that replicates colonial histories [59, 82]. This concern, coupled with opaque data practices from proprietary AI systems [86], can cause marginalized users, who may already have distrust these technological systems [53], to avoid these systems entirely, furthering the digital divide of who uses emergent technologies [88].

Ultimately, users from marginalized groups may be placed in a position where avoiding harm may require increased self-disclosure. This creates a structural privacy asymmetry, where users may have to choose between tolerating repeated harm or surrendering additional personal data to achieve desired interactions. Even if users reveal information about themselves to these systems, they risk being stereotyped, rather than the system truly aligning with their preferences. Prior work has described this phenomena as a "personalization double bind" for marginalized users, where refusing personalization may expose them to majority-oriented defaults, while accepting personalization risks stereotyping or overfitting to group identity [91].

5 Towards Understanding Harm As User-Centered, Adaptive Processes

Taken together, we argue that harm in personalized generative AI systems cannot be adequately understood as a fixed property of model outputs, nor as a quantity that can be fully specified through existing auditing paradigms. Instead, harms in these systems emerge through ongoing interaction, shaped by personalization, user history, shifting social context, and evolving harm interpretations. As such, we argue that understanding and addressing harm in personalized generative AI systems requires reconceptualizing understanding harm as a user-centered, adaptive process. Personalization can amplify the interpretative nature of harm by exposing users to system behaviors that are indeterminate, context-dependent, and differentially interpreted. As a result, users should not merely be recipients of system behavior, but active participants in defining what constitutes harm for them and how this may shift over time.

This reconceptualization has implications for how we understand the goals of auditing. Existing auditing paradigms primarily aim to surface harms for model developers, often in service of downstream technical remediation [56, 70, 73]. Users who experience harm rarely receive immediate feedback or assurance that their experiences will shape future system behavior. In personalized systems, where harms arise through interaction itself, auditing that remains external to user experience risks missing precisely the harms that are most impactful. We argue that auditing should shift toward supporting users directly in shaping less harmful interactions, rather than solely producing retrospective assessments for developers.

5.1 Design Directions

We advocate for system designs that move personalized generative AI systems away from silently inferring what constitutes harm and toward infrastructures that support explicit articulation, negotiation, and revision of harm over time. Rather than designers, auditors, or models unilaterally determining what is harmful for users, systems should create space for users to express when outputs feel harmful, reductive, or misaligned, and to do so without assuming

that harm is stable or universally agreed upon, even within shared demographic groups.

One possible direction is to embed lightweight, optional mechanisms that allow users to indicate moments of harm during interaction and, if they choose, contextualize why a response was problematic. These signals would not function as ground truth labels, but as expressions of user interpretation that can directly shape future interactions. Importantly, this reframes personalization not as the system "learning the user," but as an ongoing process in which users retain agency over how system behavior adapts.

However, as we outlined in Section 4, relying solely on individual-level articulation risks reproducing asymmetries of labor, disproportionately burdening users, particularly those from marginalized groups, with the work of identifying and explaining harm. To mitigate this, individual feedback could be complemented by community-mediated signals, where users can encounter harm characterizations articulated by others with shared or adjacent experiences and choose whether they resonate. For example, users might opt into, reject, or modify shared descriptions of harmful behaviors, without being required to generate explanations from scratch.

Crucially, community signals should not be treated as authoritative or homogenizing. Instead, they should make visible disagreement, variation, and internal plurality within groups, resisting the flattening of identities into normative standards. Users should retain the ability to refuse personalization altogether, selectively adopt shared interpretations, or revise prior signals as their understanding evolves. Across both individual and community layers, participation must remain voluntary, legible, and non-punitive. Safety should not be contingent on user labor, nor should avoiding harm require excessive self-disclosure.

More broadly, these design directions suggest a shift in the purpose of auditing itself: from identifying harms for external remediation toward enabling infrastructures through which harm can be expressed and addressed situationally. While these approaches raise open questions, such as how to balance conflicting signals, prevent misuse, and avoid reinforcing dominant norms, they underscore the central idea that understanding harm in personalized generative AI requires mechanisms for more user control and harm articulation, not just improved metrics or more granular personalization.

5.2 Limitations and Open Challenges

Our design directions raises several limitations and open challenges. Surfacing harm at the interaction level requires trust. Users, particularly those from marginalized communities, may hesitate to disclose harm in systems operated by large institutions with opaque data practices or histories of extraction [53]. Harm-reporting mechanisms themselves can be misused, ignored, or weaponized, especially when power remains centralized with platform providers. Surfacing end-user harms would be most ideal in systems that are governed by communities themselves. Acknowledging this, we make recommendations that fit into the current structure of personalized systems, but advocate for longer term, structural change that consolidates power towards end-users, not just model developers. We therefore view these design directions as incremental steps, rather than a substitute for necessary structural change in the governance and development of these systems.

Furthermore, treating harm as adaptive complicates evaluation: pluralism, disagreement, and change over time resist clean measurement and raise difficult questions about accountability and governance. Technical, pluralistic alignment across different demographic groups remains an open challenge in AI systems [2, 81], and pluralistic alignment within groups is therefore a challenge as well. Our design recommendations do not address these technical challenges, and rather focus on concrete ways for users to identify and articulate harms within the constraints of personalized AI platforms. Ultimately, understanding and preventing harms in personalized generative AI systems cannot be solved through design or auditing alone. It requires sustained governance, institutional reflexivity, and meaningful power-sharing with the communities most affected by these systems.

6 Conclusion

As generative AI systems become increasingly personalized and embedded in everyday interaction, prevailing auditing approaches, static, simulated, and group-aggregated, are ill-suited to capture the harms users actually experience. We argue that many consequential harms in personalized generative AI systems arise at the level of interaction, shaped by evolving user history, interpretation, and preferences, and therefore cannot be fully specified or evaluated outside real-world use. Moreover, attempting to close this gap through deeper personalization risks imposing asymmetric burdens of labor and privacy on marginalized users. Taken together, these challenges motivate a shift from treating harm as a static property of model outputs to understanding harm as an adaptive, user- and community-centered process. Recognizing harm as something that emerges, is contested, and evolves through interaction is essential for developing auditing and design practices that more faithfully reflect lived user experience in personalized AI systems.

Acknowledgments

I am grateful to Emily Tseng and Solon Barocas for feedback.

References

[1] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.

[2] Dalia Ali, Dora Zhao, Allison Koenecke, and Orestis Papakyriakopoulos. 2025. Operationalizing Pluralistic Values in Large Language Model Alignment Reveals Trade-offs in Safety, Inclusivity, and Model Behavior. *arXiv preprint arXiv:2511.14476* (2025).

[3] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing Race and Gender Discrimination in Online Housing Markets. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 24–35. doi:10.1609/icwsm.v14i1.7276

[4] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[5] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. arXiv:2401.14462 [cs.CY] https://arxiv.org/abs/2401.14462

[6] April Bleske-Rechek, Robert O Deaner, Katie N Paulich, Michael Axelrod, Stephanus Badenhorst, Kai Nguyen, Eleni Seyoum, and Parker S Lay. 2023. In the eye of the beholder: Situational and dispositional predictors of perceiving harm in others' words. *Personality and Individual Differences* 200 (2023), 111902.

[7] Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[8] Courtney M Bonam, Vinoadharen Nair Das, Brett R Coleman, and Phia Salter. 2019. Ignoring history, denying racism: Mounting evidence for the Marley hypothesis and epistemologies of ignorance. *Social Psychological and Personality Science* 10, 2 (2019), 257–265.

[9] Ayoub Bouguettaya, Elizabeth M Stuart, and Elias Aboujaoude. 2025. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. *npj Digital Medicine* 8, 1 (2025), 332.

[10] Paul J Brancaleone, Roberto U Cofresí, Hannah I Volpert-Esmond, David M Amodio, Tiffany A Ito, and Bruce D Bartholow. 2025. Within-person dynamics of attention to race and expression of race bias: a real-time test of the self-regulation of prejudice model. *Social cognitive and affective neuroscience* 20, 1 (2025), nsaf026.

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

[12] Rhitu Chatterjee. 2025. *Their Teenage Sons Died by Suicide. Now, They Are Sounding an Alarm About AI Chatbots.* NPR. https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide

[13] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[14] Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189* (2023).

[15] Patricia Hill Collins, Elaini Cristina Gonzaga da Silva, Emek Ergun, Inger Furseth, Kanisha D Bond, and Jone Martínez-Palacios. 2021. Intersectionality as critical social theory: Intersectionality as critical social theory, Patricia Hill Collins, Duke University Press, 2019. *Contemporary Political Theory* 20, 3 (2021), 690.

[16] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need.* The MIT Press.

[17] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and addressing algorithmic bias in practice. *Interactions* 25, 6 (2018), 58–63.

[18] Kimberlé Crenshaw. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*. Routledge, 23–51.

[19] Kimberlé Williams Crenshaw. 2013. Mapping the margins: Intersectionality, identity politics, and violence against women of color. In *The public nature of private violence*. Routledge, 93–118.

[20] Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937* (2024).

[21] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I. Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. arXiv:2501.01397 [cs.HC] https://arxiv.org/abs/2501.01397

[22] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[23] Joel Eapen and VS Adhithyan. 2023. Personalization and customization of llm responses. *International Journal of Research Publication and Reviews* 4, 12 (2023), 2617–2627.

[24] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 62–71.

[25] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[26] Christopher Frauenberger, Judith Good, Geraldine Fitzpatrick, and Ole Sejer Iversen. 2015. In pursuit of rigour and accountability in participatory design. *International journal of human-computer studies* 74 (2015), 93–106.

[27] Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning llm agents by learning latent preference from user edits. *Advances in Neural Information Processing Systems* 37 (2024), 136873–136896.

[28] Jiayi Geng, Howard Chen, Ryan Liu, Manoel Horta Ribeiro, Robb Willer, Graham Neubig, and Thomas L Griffiths. 2025. Accumulating Context Changes the Beliefs of Language Models. *arXiv preprint arXiv:2511.01805* (2025).

[29] Sourojit Ghosh, Nina Lutz, and Aylin Caliskan. 2024. "I Don't See Myself Represented Here at All": User Experiences of Stable Diffusion Outputs Containing Representational Harms across Gender Identities and Nationalities. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, Vol. 7. 463–475.

[30] Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. Do generative AI models output harm while representing non-Western cultures: Evidence from a community-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 476–489.

[31] Jonathan Grudin. 1988. Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*. 85–93.

[32] Kai Guo and Danling Li. 2024. Understanding EFL students' use of self-made AI chatbots as personalized writing assistance tools: A mixed methods study. *System* 124 (2024), 103362.

[33] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding how people customize, interact, and experience personas in large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

[34] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–25.

[35] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633, 8028 (2024), 147–154.

[36] Thomas Holtgraves, Rishi Sarin, Rebecca Wood, Emily Cronk, and Ana Júlia Nogueira Mourão. 2025. Interpreting Microaggressions: The Role of Discourse Context, Recipient Status, and Observers' Political Orientation. *Personality and Social Psychology Bulletin* (2025), 01461672251377809.

[37] Aspen Hopkins, Angie Boggust, and Harini Suresh. 2025. Chatbot Evaluation Is (Sometimes) Ill-Posed: Contextualization Errors in the Human-Interface-Model Pipeline. (2025).

[38] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 133–142.

[39] Young J Juhn, Euijung Ryu, Chung-Il Wi, Katherine S King, Momin Malik, Santiago Romero-Brufau, Chunhua Weng, Sunghwan Sohn, Richard R Sharp, and John D Halamka. 2022. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. *Journal of the American Medical Informatics Association* 29, 7 (2022), 1142–1151.

[40] Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474* (2023).

[41] Tae Soo Kim, Yoonjoo Lee, Yoonah Park, Jiho Kim, Young-Ho Kim, and Juho Kim. 2025. CUPID: Evaluating Personalized and Contextualized Alignment of LLMs from Interactions. *arXiv preprint arXiv:2508.01674* (2025).

[42] Md Kowsher, Ritesh Panditi, Nusrat Jahan Prottasha, Prakash Bhat, Anupam Kumar Bairagi, and Mohammad Shamsul Arefin. 2024. Token trails: Navigating contextual depths in conversational ai with chatllm. In *International Conference on Applications of Natural Language to Information Systems*. Springer, 56–67.

[43] Sabina Lacmanovic and Marinko Skare. 2025. Artificial intelligence bias auditing–current approaches, challenges and lessons from practice. *Review of Accounting and Finance* ahead-of-print (2025).

[44] André Alexus Lewis et al. 2025. Unpacking Cultural Bias in AI Language Learning Tools: An Analysis of Impacts and Strategies for Inclusion in Diverse Educational Settings. *International Journal of Research and Innovation in Social Science* 9, 1 (2025), 1878–1892.

[45] Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. Hello again! llm-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 5259–5276.

[46] Serene Lim and María Pérez-Ortiz. 2024. The african woman is rhythmic and soulful: An investigation of implicit biases in llm open-ended text generation. *arXiv preprint arXiv:2407.01270* (2024).

[47] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. 2012. Recommender systems. *Physics reports* 519, 1 (2012), 1–49.

[48] Reuben Luera, Ryan Rossi, Franck Dernoncourt, Alexa Siu, Sungchul Kim, Tong Yu, Ruiyi Zhang, Xiang Chen, Nedim Lipka, Zhehao Zhang, et al. 2025. Personalizing Data Delivery: Investigating User Characteristics and Enhancing LLM Predictions. In *Companion Proceedings of the ACM on Web Conference 2025*. 1167–1171.

[49] Sarah MacCarthy, Laura M Bogart, Frank H Galvan, and David W Pantalone. 2021. Inter-group and intraminority-group discrimination experiences and the coping responses of Latino sexual minority men living with HIV. *Annals of LGBTQ public and population health* 2, 1 (2021), 1.

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

[51] Mónica Melero Lázaro, Francisco José García Ull, et al. 2023. Gender stereotypes in AI-generated images. *El Profesional de la información* 32, 5 (2023).

[52] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344. doi:10.1561/1100000083

[53] Manas Mhasakar, Rachel Baker-Ramos, Ben Carter, Evyn-Bree Helekahi-Kaiwi, and Josiah Hester. 2025. "I Would Never Trust Anything Western": Kumu (Educator) Perspectives on Use of LLMs for Culturally Revitalizing CS Education in Hawaiian Schools. arXiv:2501.17942 [cs.CY] https://arxiv.org/abs/2501.17942

[54] Nusrat Jahan Mim, Dipannita Nandi, Sadaf Sumyia Khan, Arundhuti Dey, and Syed Ishtiaque Ahmed. 2024. In-between visuals and visible: The impacts of text-to-image generative ai tools on digital image-making practices in the global south. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[55] Mirja Mittermaier, Marium M Raza, and Joseph C Kvedar. 2023. Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digital Medicine* 6, 1 (2023), 113.

[56] Jakob Mökander. 2023. Auditing of AI: Legal, ethical and technical approaches. *Digital Society* 2, 3 (2023), 49.

[57] Mathieu Molina and Patrick Loiseau. 2022. Bounding and approximating intersectional fairness through marginal fairness. *Advances in Neural Information Processing Systems* 35 (2022), 16796–16807.

[58] Ariel J Mosley and Monica Biernat. 2025. Social identity and the psychology of cultural appropriation. *Handbook of Social Identity Research* (2025), 326–344.

[59] Sukanya Kannan Moudgalya and Sai Swaminathan. 2024. Toward Data Sovereignty: Justice-oriented and Community-based AI Education. In *Proceedings of the 2024 on RESPECT Annual Conference* (Atlanta, GA, USA) *(RESPECT 2024)*. Association for Computing Machinery, New York, NY, USA, 94–99. doi:10.1145/3653666.3656107

[60] Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*. 198–219.

[61] Sundaram Narayanan, Lavanya Koppaka, Narasimha Edala, Don Loritz, and Raymond Daley. 2004. Adaptive interface for personalizing information seeking. *CyberPsychology & Behavior* 7, 6 (2004), 683–688.

[62] Alexander Tobias Neumann, Yue Yin, Sulayman Sowe, Stefan Decker, and Matthias Jarke. 2024. An llm-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education* (2024).

[63] BBC News. 2026. *Changes to Grok's AI safeguards and concerns over generated sexualised imagery*. BBC News. https://www.bbc.com/news/articles/cvg1mzlryxeo

[64] Quynh Nguyen. 2022. Evaluation in participatory design–The whys and the nots. In *Proceedings of the Participatory Design Conference 2022-Volume 2*. 161–166.

[65] Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2025. User-llm: Efficient llm contextualization with user embeddings. In *Companion Proceedings of the ACM on Web Conference 2025*. 1219–1223.

[66] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[67] Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 496–511.

[68] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. 2024. User Modeling and User Profiling: A Comprehensive Survey. arXiv:2402.09660 [cs.AI] https://arxiv.org/abs/2402.09660

[69] Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The Case for "Thick Evaluations" of Cultural Representation in AI. arXiv:2503.19075 [cs.CY] https://arxiv.org/abs/2503.19075

[70] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[71] Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430* (2024).

[72] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2010. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.

[73] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2019. Aequitas: A Bias and Fairness Audit Toolkit. arXiv:1811.05577 [cs.LG] https://arxiv.org/abs/1811.05577

[74] Alejandro Salinas, Amit Haim, and Julian Nyarko. 2024. What's in a name? Auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875* (2024).

[75] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22, 2014 (2014), 4349–4357.

[76] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?. In *International Conference on Machine Learning*. PMLR, 29971–30004.

[77] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine* 27, 12 (2021), 2176–2182.

[78] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[79] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Chenglei Si, et al. 2024. Position: Towards Bidirectional Human-AI Alignment. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*.

[80] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 6 pages. doi:10.1145/3551624.3555285

[81] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070* (2024).

[82] Isabella Spano and Yuxiao Zhang. 2025. Indigenous data sovereignty in intangible cultural heritage governance: A complementary approach to public–private partnerships. *International Journal of Cultural Property* (2025), 1–27. doi:10.1017/S0940739125100064

[83] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[84] Vishesh Thakur. 2023. Unveiling gender bias in terms of profession across LLMs: Analyzing and addressing sociological implications. *arXiv preprint arXiv:2307.09162* (2023).

[85] Victor Udoewa. 2022. An introduction to radical participatory design: decolonising participatory design processes. *Design Science* 8 (2022), e31.

[86] Bram Vaassen. 2022. AI, opacity, and personal autonomy. *Philosophy & Technology* 35, 4 (2022), 88.

[87] Keyon Vafa, Sarah Bentley, Jon Kleinberg, and Sendhil Mullainathan. 2025. What's Producible May Not Be Reachable: Measuring the Steerability of Generative Models. doi:10.48550/arXiv.2503.17482 arXiv:2503.17482 [cs].

[88] Jan Van Dijk and Kenneth Hacker. 2003. The digital divide as a complex and dynamic phenomenon. *The information society* 19, 4 (2003), 315–326.

[89] Angelina Wang. 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*. 485–497.

[90] Angelina Wang, Xuechunzi Bai, Solon Barocas, and Su Lin Blodgett. 2025. Measuring Machine Learning Harms from Stereotypes Requires Understanding Who Is Harmed by Which Errors in What Ways. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 746–762.

[91] Angelina Wang, Erin Beeghly, Sanmi Koyejo, and Daniel E. Ho. 2025. Personalization Double Binds: When User Preferences Meet Group-Based Chatbot Behaviors. *arXiv preprint* (2025).

[92] Angelina Wang, Daniel E. Ho, and Sanmi Koyejo. 2025. The inadequacy of offline large language model evaluations: A need to account for personalization in model behavior. *Patterns* 6, 12 (2025), 101397. doi:10.1016/j.patter.2025.101397

[93] Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence* (2025), 1–12.

[94] Athena Wen, Tanush Patil, Ansh Saxena, Yicheng Fu, Sean O'Brien, and Kevin Zhu. 2025. FAIRE: Assessing Racial and Gender Bias in AI-Driven Resume Evaluations. *arXiv preprint arXiv:2504.01420* (2025).

[95] Theoni Whyman, Karen Adams, Adrian Carter, and Laura Jobson. 2021. Lateral violence in indigenous peoples. *Australian Psychologist* 56, 1 (2021), 1–14.

[96] Theoni Whyman, Cammi Murrup-Stewart, Adrian Carter, Uncle Michael Young, and Laura Jobson. 2022. Ngarratja Kulpaana: Talking together about the impacts of lateral violence on aboriginal social and emotional well-being and identity. *Cultural diversity & ethnic minority psychology* 28, 2 (2022), 290.

[97] Kyra Wilson and Aylin Caliskan. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 1578–1590.

[98] Sophia Worth, Ben Snaith, Arunav Das, Gefion Thuermer, and Elena Simperl. 2024. Ai data transparency: an exploration through the lens of ai incidents. *arXiv preprint arXiv:2409.03307* (2024).

[99] Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025. Personalized generation in large model era: A survey. *arXiv preprint arXiv:2503.02614* (2025).
[100] Yiran Yang. 2025. Racial bias in AI-generated images. *AI & SOCIETY* (2025), 1–13.
[101] Iris Marion Young. 2002. *Inclusion and democracy*. OUP Oxford.
[102] Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology* 21, 2 (2019), 89–103.
[103] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2023. Coding inequity: assessing GPT-4's potential for perpetuating racial and gender biases in healthcare. *MedRxiv* (2023), 2023–07.
[104] Lili Zhang, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024. Partiality and misconception: Investigating cultural representativeness in text-to-image models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.
[105] Kaitlyn Zhou, Kristina Gligorić, Myra Cheng, Michelle S Lam, Vyoma Raman, Boluwatife Aminu, Caeley Woo, Michael Brockman, Hannah Cha, and Dan Jurafsky. 2025. Attention to Non-Adopters. *arXiv preprint arXiv:2510.15951* (2025).
[106] Leying Zou and Warut Khern-am nuai. 2023. AI and housing discrimination: The case of mortgage applications. *AI and Ethics* 3, 4 (2023), 1271–1281.

Received 20 February 2026; accepted 2 March 2026