

Shaping the Future of Generative AI for Black Communities: A Frame Analysis of Public Discourse and Empirical Scholarly Research

Angela D. R. Smith¹, Gabriella Thompson², Christopher L. Dancy³
Mark Diaz⁴, Seyi Olojo⁵, Christina N. Harrington⁴

¹University of Rochester

²University of Texas at Austin

³The Pennsylvania State University

⁴Google Research

⁵University of California
Berkeley

a.smith@rochester.edu

Abstract

As generative AI (genAI) systems become embedded in education, employment, healthcare, and creative industries, the impact and engagement among marginalized groups have become both a widespread discourse and a focus in scholarly research. As a starting point, we examine public discourse and empirical research to explore the impact of genAI systems on Black communities. We conducted a systematic literature review (SLR) of 91 empirical papers alongside a media discourse frame analysis of 28 public resources, applying Entman's framing theory to map how each corpus defines problems, attributes causes, and proposes treatments. Our SLR reveals that scholarly research concentrates heavily on technical bias detection, reducing Blackness to measurable variables rather than engaging with cultural practices, structural conditions, or Black knowledge systems. Our frame analysis reveals that public discourse attributes genAI-related harm to historical and systemic forces, while scholarly research stops its causal accounts at the dataset and its treatment recommendations at technical reform. We demonstrate that this misalignment is structurally produced: anti-Blackness operates simultaneously across both registers, generating a shared evacuation of Black epistemic agency. We argue for frame analysis as an AI ethics methodology capable of surfacing what technical evaluation forecloses.

Introduction

Public discourse shapes how societies understand, adopt, and regulate emerging technologies. As generative AI (genAI) systems become increasingly embedded in education, employment, healthcare, and creative industries, narratives circulating in media outlets, policy briefs, and opinion pieces play a crucial role in signaling anticipated harms, assigning responsibility, and proposing policy and technical interventions. For historically marginalized communities, particularly Black Americans who have borne disproportionate harms from algorithmic systems ranging from biased facial recognition to discriminatory hiring tools (Noble 2018; Benjamin 2019), understanding how public discourse frames genAI's potential impacts is essential for informing

research agendas, policy priorities, and community organizing efforts. Similarly, empirical research and the translational impact observed in research papers have begun to directly influence the policy, design, development, and use of genAI tools (Septiandri, Constantinides, and Quercia 2024). Such scientific explorations are especially pivotal to addressing existing and anticipated harms, underscoring the value of the knowledge cultivated by scholarly research.

This surfaces a critical question that remains underexplored: How do the concerns and framings present in public discourse about genAI and Black communities align with—or diverge from—the priorities of research that seeks to directly study and engage with these communities? As this work evolves, it is also essential to interrogate conceptions of Blackness and *how* Black communities are engaged and the ways Black culture and identity are portrayed to inform AI development. Considering such interactions requires us to interrogate the extractive nature of AI research and development (Network 2025), and move beyond the oppressive nature of “inherently biased algorithms”. If public discourse emphasizes the importance of community-driven approaches to identifying and mitigating structural harms such as cultural erasure (i.e., when AI corrects for bias by censoring parts of Black history (Small 2023)), scholarly research may miss the mark of potential impact by focusing predominantly on automating the evaluation of technical metrics such as bias detection. This misalignment may indicate that research is failing to address the issues that matter most to affected communities and the broader public, while also flattening Blackness to merely be categorical dimensions that typically do not capture the full human condition of this community.

This paper investigates this potential gap through a frame analysis of public discourse concerning genAI and Black communities. Drawing on Entman's (1993) widely-used framing theory, we systematically analyze media resources—including news articles, opinion pieces, and policy briefs—to identify how public narratives define problems, attribute causes, render moral judgments, and propose treatments related to genAI's impact on Black Americans. We complement this analysis with a systematic literature review

(SLR) of empirical literature from research institutions published at computing venues to characterize how research has approached genAI and Black communities, providing context for assessing alignment between public discourse and scholarly priorities.

Our analysis reveals significant divergences between public discourse and empirical scholarly research. Using framing theory signals and emphasizes several dimensions of harm that can be attributed to systemic bias and marginalization, while the SLR suggests a heavy research focus on technical bias detection. We offer three contributions to the AIES community. First, we identify and characterize the dominant frames shaping public narratives about genAI's anticipated impacts on Black communities. Second, we demonstrate the misalignment between public discourse and empirical research, identifying it as not incidental but structurally produced. Third, we argue for frame analysis as an AI ethics methodology — one that captures the structural and anticipatory dimensions of harm that individual-focused technical evaluation forecloses, and that makes visible what research epistemology is structured to exclude. In doing so, we respond directly to AIES's remit of interrogating the societal dimensions of AI: frame analysis extends the community's methodological repertoire beyond system-level evaluation toward the discursive and structural conditions under which AI harms are defined, contested, and governed.

We argue that attending to public discourse offers valuable insights for AI ethics scholarship. Public frames reveal what harms are anticipated before they fully materialize, what causal explanations resonate with broader audiences, and what interventions are seen as viable—information that can inform both research agendas and policy development. By mapping the terrain of public concern and research focus, this paper contributes to ongoing efforts to ensure that AI research addresses the issues that matter to the communities most likely to be affected by these powerful technologies.

Blackness, Racialization & Relation to GenAI

Understanding the relationship between genAI and Black communities requires situating contemporary bias within the historical processes of racialization that shape how AI systems are developed, deployed, and experienced. While Black people, communities, and practices trace origins well beyond recent forms of racialization, current racial structures significantly influence technology at micro (individual identities and interpersonal interactions), meso (organizational and institutional practices), and macro (societal structures, law, and policy) levels (Ray 2019). The logics underlying race and Blackness from Western, white perspectives, what Wynter and McKittrick (Wynter and McKittrick 2015) describe as the “standard genre of the human”, are embedded in the engineering lifecycle of AI systems (Workman and Dancy 2026), producing the specific manifestations of bias documented in empirical research.

These racialized logics have evolved from theological to biological to cultural rationales, but consistently position Blackness as inferior or deviant (Robinson 2020; Kendi 2017; Wynter 2003), operating through what Workman and Dancy (Workman and Dancy 2026) identify as “inlets”

where anti-Blackness enters AI development at every level. Within Human-Computer Interaction (HCI), this manifests through “solutionist” approaches that treat Blackness as a problem to be corrected rather than recognizing the ontological complexity of Black experience and knowledge (Cunningham et al. 2023). Colorblind perspectives (Bonilla-Silva 2015) and the fluidity of racial categories further obscure the persistence of these logics, making structural intervention difficult even as Black communities continue to imagine with and beyond racialization (Bray and Harrington 2021).

Related Work

Field Goals and Motivation of GenAI

The values structuring AI development shape how systems are built and assessed, with consequences for minoritized communities. Birhane et al. (2022b) identify that AI scholarship prioritizes quantitative empiricism, efficiency, and performance over societal need and negative impacts — a pattern reflected in the field's reliance on benchmark datasets, which have been shown to value efficiency and universality over care and contextuality (Scheuerman, Hanna, and Denton 2021; Qadri et al. 2025). These field-level values extend to fairness paradigms: Hanna et al. (2020) argue that dominant approaches treat oppressed identities as interchangeable, lacking race-consciousness, while Blodgett et al. (Blodgett et al. 2020) call for researchers to account for power relations between themselves and the communities they study.

Biases and Harms in GenAI

Scholarly research has documented a range of social biases in genAI systems, from racial bias in language models (Sap et al. 2019) to gender bias in image generation (Wan et al. 2024), generating frameworks for identifying and mitigating model bias (Díaz et al. 2024). Shelby et al. (Shelby et al. 2023) offer the most comprehensive taxonomy of sociotechnical algorithmic harms — representational, allocative, quality of service, interpersonal, and social system — which we adopt in our analysis. Research specifically focused on Black communities has identified failures to understand African American communication (Brewer, Harrington, and Heldreth 2023), inadequate support for Black cultural contexts (Olojo et al. 2025), and reproduction of cultural stereotypes in genAI outputs (Hofmann et al. 2024). Despite the value of this work, researchers have called for new evaluation methods commensurate with genAI's scale and scope, including methods rooted in HCI (Weidinger et al. 2023).

Bridging Black Communities and AI Development

Public perceptions of AI influence how systems are developed and deployed (Lima, Grgić-Hlača, and Cha 2023; Liu et al. 2024), and these perceptions vary across race, class, and gender (Kelley et al. 2021). Black communities hold complex, often critical attitudes toward AI — particularly in healthcare contexts (Lee and Rich 2021; Parker, Vardoulakis, and Harrington 2026) — and report

pessimism about algorithmic fairness and bias in online platforms (Harris et al. 2023). These perceptions have motivated calls for responsible inclusion of Black communities in AI development, including feminist-reflexive data curation (Leavy, Siapera, and O’Sullivan 2021), critical race theory for AI fairness (Hanna et al. 2020), and CRT-grounded HCI (Ogbonnaya-Ogburu et al. 2020; Smith et al. 2020). Participatory approaches have demonstrated particular value: co-design with Black adults for mental health agents (O’Leary et al. 2025), speculative design with Black youth for justice-oriented AI principles (Tanksley et al. 2025), and community workshops for equitable speech technology (Brewer, Harrington, and Heldreth 2023). Yet scholars note that participation alone is insufficient without intentional structural commitment (Sloane et al. 2022; Birhane et al. 2022a). Our work contributes to this area by examining the parallel terrain of public discourse and empirical research to map where emphases, priorities, and concerns align and where they diverge.

Methods

To map this landscape, we employed a dual-methodological approach combining a systematic literature review (SLR) with a frame analysis of discourse found in public media and reports. Contrasting public concerns and priorities with focus areas dominating empirical research helps to reveal misalignments that may indicate where research is failing to address community-identified needs and anticipatory harms. We detail our process below (see also Figure 3).

Systematic Literature Review

Data Collection and Corpus Creation To capture the landscape of research on genAI conducted with Black communities, we conducted a systematic literature review of full research papers published in major computing and AI databases between January 2010 and December 2025. We sought to assess the focus and priorities of existing research, to examine how Black communities were engaged in the ideation, evaluation, and testing of genAI tools, and to identify existing research on their perceptions, attitudes, and opinions of genAI (Haimson et al. 2025). We focused on publications from the ACM (encompassing HCI venues including CHI, CSCW, DIS, FAccT, and TOCHI) and IEEE digital libraries, and specialized AI venues, such as ACL, EAAMO, and NeurIPS.^{1,2} We also searched ArXiv, a venue where several prominent AI papers have been published to date. While other venues in the social sciences and human-

¹We initially included AAAI in our database search; however, in trying to repeat some of the searches, we found the search system itself produced inconsistent results across repeated identical queries, undermining the reproducibility our protocol required. As such, we removed those papers from our corpus and note this as a coverage limitation, given AAAI’s prominence as an AI/ML venue.

²Scopus was excluded due to its substantial overlaps with the ACM and IEEE digital libraries. Google Scholar was considered unsuitable as a principal search system for a systematic review due to its non-transparent ranking, personalized results, and unstable retrieval (Gusenbauer and Haddaway 2020).

ities may have been relevant to the conversation of understanding the reception of genAI among various historically marginalized communities, we prioritized papers published in the computing field.

Three members of the research team defined search terms within two categories: “*solo/group racial identity terms*” and “*generative AI terms*”. *Solo racial identity terms* included ‘Black communities’, ‘Black populations’, or ‘Black Americans’. *Group racial identity terms* included terms like ‘people of color’, ‘marginalized/minoritized communities’, or ‘vulnerable populations’. Our search also incorporated all possible generative AI terms including ‘large language models’, ‘text-to-text’, ‘ChatGPT’, ‘AI image generation’, and ‘prompt engineering’ and derivations of these terms.

We then searched using a combination of terms from each category, for example: ‘Black people’ + ‘generative AI’, or ‘African American’ + ‘text-to-image’. We filtered research articles with the following inclusion criteria: 1) Must focus on genAI (images, text, and/or videos); 2) Must be a full peer-reviewed paper published between 2010-2025; 3) Be published in English; 4) Must be empirical research where Black population is the focus of a qualitative study (interviewing, focus groups, participatory design), quantitative study (testing and evaluation of systems, surveys), or computational techniques (dataset and model evaluation frameworks). Works in progress, conference extended abstracts, and poster submissions were omitted.

All authors iterated on a survey coding instrument and survey items to ensure mutual understanding during coding. We chose five papers to test the coding instrument for reliability and validity among researchers. Once the coding instrument was finalized, the remaining papers in the larger corpus were split and coded by one of four researchers.

Our search returned 223 papers across the five databases. Four members of the research team completed a title and abstract search to remove papers that did not fit the purpose of this review (n=71), and any duplicate papers (n=61) that may have returned over multiple databases. We then removed 18 papers that did not focus on Black communities but marginalized groups more broadly, leaving 91 papers for coding and analysis (19 IEEE, 41 ArXiv, 13 ACM, 18 ACL) (Figure 3).

Analysis Our coding instrument was used to analytically process dimensions and concepts and deductively extract data from 91 research articles. Papers were coded for year of publication, venue, search terms present in the paper, type of research study, type of genAI, intended contribution, domain or context of genAI application, whether the paper focused on an artifact or dataset, whether the artifact or dataset was evaluated, and forms of engagement. We also assessed whether the authors defined Blackness or other terms of identity. We used an aggregate analysis of closed-ended questions to describe trends in publication timelines and venues, the frequency of paper topics by type of genAI system, and research types. Open-ended questions such as ‘Contribution’ and ‘Purpose/Goal’ were analyzed thematically. After our initial coding of the paper corpus, we noticed that our papers largely centered harms and harm mitigation.

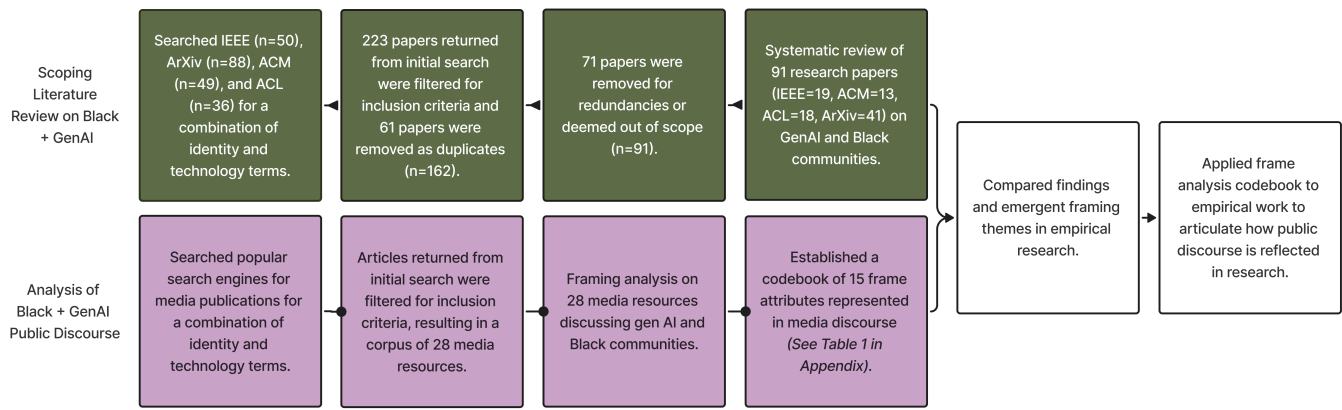


Figure 1: Systematic Review and Framing Analysis flow diagram.

As such, we recoded our corpus using the frame attributes, notably Shelby et al. (2023)’s sociotechnical harms, which helped us articulate the harms present in the papers.

Frame Analysis

We use frame analysis to examine how public discourse shapes conversations about the impact of genAI technology production on Black communities. Our frames, organizing principles that structure the discourse about the social impact of AI, give us the opportunity to understand how the public both predicts and reacts to the inclusion of genAI within numerous industries, including finance, healthcare, and the service sector. This approach also illustrates the driving factors influencing public discourse about genAI, specifically for and within marginalized communities.

Data Collection and Corpus Creation We used keywords such as: “*Generative AI*”, “*African American*”, “*Black*”, “*Black American*”, and “*genAI*” to source popular media to analyze discourse on Black American communities and their relationship to genAI (Baumer et al. 2017; Zade et al. 2024). Three authors helped sample various sources, including the popular press, op-eds, investigative journalism, blogs, and white papers from research organizations and think tanks. Our final corpus of text includes 28 articles. Note that our inclusion logic differs across the two corpora by design: the SLR characterizes the peer-reviewed empirical research record, and so excluded non-archival formats such as extended abstracts and posters, whereas the frame analysis characterizes public discourse, of which white papers and think tank reports are a constitutive genre alongside journalism and opinion writing.

Analysis Extending prior work (Zhang et al. 2025; Zade et al. 2024), we leverage Entman’s description of framing as a way to “*select some aspects of a perceived reality and make them more salient in a communicating text*” (Entman 1993). Two authors conducted an initial thematic analysis of the 28 articles to code frame characteristics, such as the extent to which positive or negative descriptions of genAI were included in the text (Van Hulst et al. 2025).

For the second round of analysis, we coded each source

according to Entman’s four functions of a frame: “*to promote a particular problem definition; causal interpretation; moral evaluation; and/or treatment recommendation*” (Entman 1993). Our guide for this portion of analysis stems from Matthes and Kohring’s framing methodology, which understands frames as “*a certain pattern in a given text that is composed of several elements. These elements are not words but previously defined components or devices of frames*” and suggest treating each of Entman’s frame elements as composed of several analytical devices, which we term “Frame attributes” (Matthes and Kohring 2008).

Coding Frame Attributes We drew upon the insights identified in the initial round of coding to create our frame attributes for each of Entman’s frame elements (see Tables 2, 3, 4, 5), excluding the subcategories for ‘Problem definition’, which were derived from Shelby et al. (2023)’s taxonomy of sociotechnical harms from algorithmic systems. We leveraged this taxonomy after recognizing that the main problems identified in our first round of analysis aligned with the harms defined in Shelby et al. (2023), and we then used it to deductively code the corpus. To determine the frame attributes for subsequent frame elements, we first revisited the codes from the first round of thematic analysis that captured framing themes from the corpus. We used these categories as guidelines to refine and create new attributes that best captured Entman’s four frame elements.

Problem Definition. Our initial analysis of the corpus revealed that media resources scrutinized genAI’s impact on Black communities and often focused on the potential for genAI to produce algorithmic harm. Recognizing this theme, we leveraged Shelby et al. (2023)’s taxonomy for sociotechnical harms of algorithmic systems as our frame attributes for Entman’s first frame function, *Problem definition*, which occurs when an author describes “*what a causal agent is doing with what costs and benefits*” (Entman 1993). We identified five sociotechnical harms that capture the costs of genAI’s use on Black communities: social system harm, representational harm, allocative harm, quality of service harm, and interpersonal harm.

Causal Interpretation. Here, we elaborate on Entman’s second frame function, causal interpretation, defined as how

authors “*identify the forces creating the problem*” (Entman 1993). We consider Entman’s *forces* to be actors (i.e., individuals, communities, and institutions) with power. We recognized how authors would attribute the harms of genAI to actors that could be categorized under two primary causes: *Historical bias* and *Systemic bias*. We consider each cause as a frame attribute for the *Causal interpretation* frame element. Our understanding of historical bias considers forces that have historically marginalized communities, often systemic or structural, and are typically evidenced from what we’ve seen in the past. For example, a machine learning model may output gendered occupations due to training data that reflected societal biases at the time of collection (Suresh and Guttag 2021). We understand systemic bias as the oppression of Black people through “*policies, customs, or behaviors that are part of the culture or structure of an organization*” (Mehrabi et al. 2021). These codes often overlapped due to the intersectional nature of historical and systemic oppression; we differentiated them to capture the range of causal agents described by the media.

Moral Judgment. The third function of Entman’s framing theory is moral judgment, in which authors “*evaluate causal agents and their effects*” (Entman 1993). We determined the subject’s moral condition by coding the benefits and risks associated with it (Matthes and Kohring 2008). We applied a similar analytical method by coding the resources according to whether they addressed genAI’s risks and benefits. For the *Moral judgment* frame element, we identified two frame attributes: *GenAI is harmful* and *GenAI can be both harmful and beneficial*. Because these frame attributes were holistic reflections of a resource, they could not be double-coded. If an article discussed only the negative implications of genAI, with no mention of its benefits, we coded it as *GenAI is harmful*.

Treatment Recommendation. Following their description of the problem, authors often suggested methods to avoid or alleviate the negative effects of genAI. Proposing a treatment recommendation is the last of Entman’s framing functions, which occurs when authors “*offer and justify treatments for the problems and predict their likely effects*” (Entman 1993). We coded *Treatment recommendation* by identifying any strategies or calls to action in the text. Through our analysis, we identified five recurring strategies which capture the treatments suggested in media discourse: Expand access to genAI, Resist genAI, Regulate genAI, Devise strategies to mitigate harm of genAI, and Reform genAI.

Findings

We present our findings in two parts before putting them in conversation. First, we report descriptive results from our systematic literature review, characterizing the research landscape, methodological trends, and community engagement patterns across the 91-paper corpus. Second, we report descriptive results from our frame analysis, summarizing how Entman’s four frame elements appeared across the 28 media resources. Having established what each corpus contains on its own terms, we then organize our findings around three of Entman’s four frame functions, *problem definition*, *causal interpretation*, and *treatment recommenda-*

tion, reading patterns across both corpora through Dancy and Saucier (2022)’s onto-epistemological framework to surface not only where public discourse and scholarly research diverge, but what their convergences reveal.

Systematic Literature Review Results

Trends in Types of Empirical Research While our search spanned 2010 to 2025, our search showed relevant papers published only between 2021 and 2025, with the largest numbers published in the last three years: 2023 (n=9), 2024 (n=16), and 2025 (n=62).

Assessing the domain and context of application of genAI, we found that an overwhelming number of papers (n=54) papers focused on addressing racial and gender bias in text-to-text and text-to-image models. When assessing the specific domain topic areas that these research papers focused on, three papers focused on Economics (e.g., Ren et al. 2024; Lakkaraju et al. 2023); ten papers focused on Health (e.g., Wang and Zhai 2024; Bautista et al. 2023; Zhang et al. 2023); six papers reported on assessing culture (e.g., Xu 2023; Yucer et al. 2020); seven papers looked at Education (e.g., Graves et al. 2024); four papers were coded for ‘Social justice/Activism’ (e.g., Basoah et al. 2025b; Tanksley et al. 2025); and two papers looked at Archival Practices (e.g., Osti and Roke 2024; Croskey et al. 2025a). Several papers (n=7) focused on linguistic equity (e.g., Kim et al. 2025; Kempermann et al. 2026), most of which looked specifically at comparisons of African American Vernacular English (AAVE) to standard English.

Methods and Technological Focus Across the papers reviewed, results show that 52 were model evaluations, 25 were the development of a benchmark or dataset, and 12 were the development of an artifact or system (Figure 3). Four papers that focused on developing a dataset utilized surveys to inform these datasets (e.g., Sandoval et al. 2025; Bosco et al. 2025). Among the papers that focused on evaluations and assessments of datasets, four papers utilized or reported human annotation by consulting a subset of Black communities as AAVE experts to verify accuracy in translation of a dataset evaluation (Gupta et al. 2024; Okpala and Cheng 2025; Deas et al. 2023; Sandoval et al. 2025). While the main focus of our project’s systematic literature review methodology was to assess how Black communities were engaged in genAI research, only four papers engaged Black communities in needs finding for genAI systems such as interviewing (Solyst et al. 2025; Tanksley et al. 2025; Basoah et al. 2025b; Bosco et al. 2025).

Looking at the types of genAI across our corpus, we found that 16 papers focused on a specific named model or system (e.g., ChatGPT, StyleGAN2, visual language model), 18 papers focused on text-to-text models, 18 papers focused on text-to-image papers, four papers focused on text-to-media (e.g., audio or video), and 1 paper reported on image-to-image (Yucer et al. 2020).

Communities of Focus and Relationality to Blackness When examining how Black communities were mentioned in the research, 28 published papers featured “Black and/or

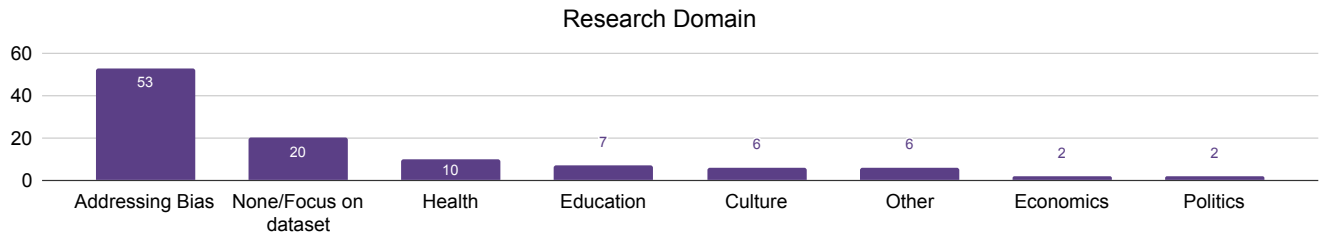


Figure 2: Distribution of domain topics among reviewed papers.

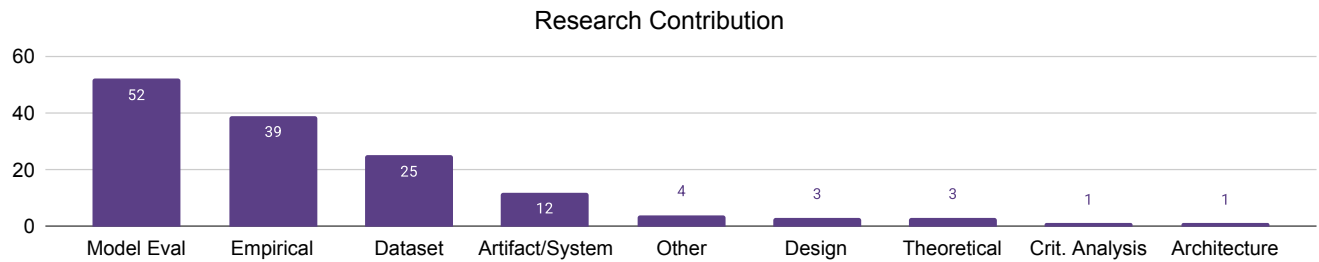


Figure 3: Distribution of paper contribution types among reviewed papers

African American” communities as the sole focus. 38 papers reported Black communities as a subset or comparison to other racial groups and communities (e.g., Zhou et al. 2024; Nezhad, Bandyapadhyay, and Agrawal 2024; Devinney, Björklund, and Björklund 2024; Guo and Caliskan 2021; Slaughter et al. 2023; Ren et al. 2024; Gupta et al. 2023; Lee et al. 2024; Xu 2023; Yucer et al. 2020; Howard et al. 2024; Wu et al. 2024), and four as a part of “People of Color (PoC)” (e.g., Kuck 2023; Huang et al. 2025). A majority of papers used datasets incorporating features of Black Americans to test models. In terms of relationality to Blackness, many papers reduced the construct of Blackness to features of skin tone (e.g., (Wilson, Ghosh, and Caliskan 2025a)) or dialect, with a majority of papers looking at the comparison of these features to other identities considered “mainstream” (e.g., focusing on the validation of AAVE language- Graves et al. 2024; Deas et al. 2023; Gupta et al. 2024). We found that only two papers included any framing of Blackness or its relationality to technology that was not categorical features. Both papers note the variation that exists in Black identity as experienced by those across the diaspora both within the United States and internationally (Basoah et al. 2025b; Croskey et al. 2025a), with the latter paper noting their definition of Blackness to encompass “the political identity of those who consciously act in solidarity with the people and traditions of this global community—in tandem with their Black ethnic and racial identities.” Among our corpus, we found that only six papers pushed for more participation of Black communities in AI development and research (e.g., Graves et al. 2024; Dossou and Aïdasso 2025; Dildora et al. 2025; Bosco et al. 2025). Some papers, while not directly engaging members of the Black community, also advocate for more participatory en-

gagement to promote co-learning, empowerment (Dossou and Aïdasso 2025), and community-centered archival practices (Croskey et al. 2025a).

Trends in Research Contributions We analyzed the various goals and empirical claims as stated in the authors’ goals and contribution statements. Many papers made straightforward claims of testing genAI models (e.g., Wang and Zhai 2024; Shankar et al. 2024; Gupta et al. 2024; Lee and Jeon 2024)), while papers like Devinney, Björklund, and Björklund (2024) state a contribution of “*investigating stereotypes, norms, and identity narratives across models*”. Others, such as Wang and Zhai (2024), report multiple contributions, including establishing an unbiased dataset, developing a model, evaluating AI-generated images, and testing the model. Figure 3 details the types of research contributions made across papers in our corpus.

Most papers referred to bias as a potential harm that disproportionately impacts Black communities or, more broadly, communities of darker skin tones (e.g., Lee and Jeon 2024; Xu 2023; Yucer et al. 2020). Papers that addressed stereotypes, norms, and identity narratives across models assessed how LLMs replicate societal stereotypes. Wu et al. (2024), for example, presents stereotype classifiers to reduce the perpetuation of societal stereotypes. Lee and Jeon (2024) explores the impact of skin tone variation across models, finding that more homogeneous stories are produced about darker skin tones, particularly among women, and highlights the need to address intersectional bias. Overwhelmingly, the majority of papers focused on anticipating various types and levels of bias and harm.

Source type	Sociotechnical harm					Treatment Recommendation				
	Representational	Allocative	Social System	Interpersonal	Qual. of Service	Expand Access	Resist	Regulate	Devise	Reform
<i>Media (28)</i>	12	8	21	12	11	6	6	13	3	12
<i>SLR (91)</i>	62	24	36	40	36	5	3	5	24	78

Table 1: The number of media sources and systematic literature review (SLR) sources expressing each frame.

Framing Results

Across the 28 media resources, causal attribution was the most consistent dimension of framing: [25] of [28] sources identified systemic bias as a causal force, and [21] of [28] identified historical bias — with over half citing both simultaneously, reflecting how the corpus treats these as structurally intertwined rather than distinct explanations. By contrast, moral judgment divided the corpus more sharply. The two resources coded as framing genAI as purely beneficial — both from Black Enterprise, covering Amazon’s AI initiatives for Black entrepreneurs and Robert Smith’s investment in HBCUs (Abdur-Rahman 2025; Enterprise 2025) — stand as notable outliers: the only sources in the corpus oriented primarily toward opportunity rather than harm. The remaining resources were split between framing genAI as harmful only and as both harmful and beneficial, with the former concentrated in sources addressing cultural representation and the latter in sources addressing economic and educational impacts, where the possibility of equitable access remained open.

Treatment recommendations further reveal the corpus’s structural orientation. Regulation was the most frequently recommended treatment [13] of [28], followed closely by reform of AI development [12] of [28]. Notably, the two most recent sources on data center infrastructure and environmental harms to Black communities — which contributed substantially to the prevalence of social system harm coding — offered no treatment recommendations at all, identifying the problem without proposing intervention. This pattern suggests that public discourse on the environmental dimensions of genAI harm remains in an early diagnostic phase, naming the issue without yet having a remedial vocabulary to match.

Across the 28 media resources and 91 empirical papers, our frame analysis identified patterns of alignment and divergence in how each corpus defined problems, attributed causes, rendered moral judgments, and proposed treatments. Table 1 summarizes the distribution of frame attributes across both corpora, and Tables 2, 3, 4, 5 provide the full breakdown of how individual sources map to each attribute. Having established what each corpus contains on its own terms, the descriptive patterns documented above — the media corpus’s structural causal orientation, the SLR corpus’s concentration on technical reform, and the shared attention to representational harm across both registers — set the conditions for a deeper analytical question: not simply where these two corpora diverge, but what their convergences and silences reveal.

Problem Definition: Representational Harm as a Shared Epistemic Frame

Although social system harm became the most frequently coded problem definition in the media corpus, a surge driven in part by the concentration of 2025 publications focused on environmental harms of generative AI alongside longstanding concerns about structural inequality, representational harm remains the organizing frame for this analysis because it is the only problem definition that appears with comparable force across both corpora, making it the most productive site for examining what that convergence reveals about how anti-Blackness structures knowledge production in public and research discourse alike. Representational harm appeared in [12] of [28] media resources and [62] of [91] research papers. On the surface, this convergence might appear to validate representational harm as the central problem framing for understanding genAI’s impact on Black communities, evidence that public discourse and empirical research are, at minimum, looking at the same issue. We argue instead that this convergence is analytically significant, and that reading it through the onto-epistemological framework of Dancy and Saucier (2022) reveals it as a symptom of anti-Blackness structuring knowledge production in both registers simultaneously, rather than confirmation of a shared and adequate problem definition.

In the media corpus, representational harm manifested through descriptive accounts of genAI outputs that erased, distorted, or stereotyped Black identity. Authors documented image generation tools that progressively lightened the skin of darker-skinned subjects with each subsequent prompt, language models that defaulted to stereotypical depictions of Black Americans, and platforms that systematically erased Black cultural content when overcorrecting for racial bias. One piece on Black artists warned that genAI may “stereotype or censor Black history and culture,” citing how tools “default to the worst stereotypes that already exist on the internet” (Small 2023). Alongside representational concerns, the media corpus also surfaced harms that the SLR corpus largely did not: social system harms appeared in [21] of [28] resources, with economic outlets emphasizing that Black Americans are “overrepresented in four sectors [...] that are likely to see a great deal of automation” and already face higher unemployment (Stewart III 2024). Educational resources identified allocative harms — the risk that AI learning tools’ documented benefits “may not reach students without resources,” expanding “achievement and opportunity gaps” (Pham et al. 2024) — and inter-

personal harms, including findings that Black students were more than twice as likely as white or Latino students to be falsely accused of using AI to write their work (Klein 2024). These harms received no parallel treatment in the SLR corpus, a divergence we return to in our discussion of causal interpretation.

In the SLR corpus, representational harm was similarly the most frequently identified problem definition, with stereotyping as the most common descriptor. Research examined stereotyping in LLM personas (Gupta et al. 2023), vision-language models that represented darker-skinned Black individuals as more homogeneous than lighter-skinned individuals (Lee et al. 2024), and image generation systems that amplified race and gender stereotypes (Chang and Wiens 2024). As with the media corpus, the dominant treatment recommendation was technical reform. The convergence across both corpora — representational harm as the problem, technical reform as the solution — reproduces what Dancy and Saucier describe as the epistemic condition in which Blackness is rendered legible primarily as an object of measurement and correction rather than as a structuring epistemology. Both public discourse and SLR research concentrate remedial effort in the register of representation — more accurate depictions, more diverse datasets, more inclusive development teams — without implicating the structural conditions that make representational harm possible and persistent.

This argument is sharpened by what is conspicuously absent from both corpora simultaneously. Neither contains frames or findings centered on Black communities as knowledge producers, technological leaders, or epistemic agents in genAI development. Among [28] media resources, no frame centered on Black-led AI innovation, community-driven design, or Black technological agency as a structural response to representational harm. Among [91] research papers (14 of which engaged Black communities directly), only [3] engaged Black community members as sources of knowledge and validation rather than as subjects of study (e.g., Tanksley et al. 2025; Dildora et al. 2025; Dorn et al. 2025), two of which positioned community epistemology as constitutive of research design. Dancy and Saucier (2022) argue that increasing representation within systems that maintain anti-Blackness as their structural condition does not dissolve that condition but transforms how it operates. The concentration on representational harm in both corpora, accompanied by the shared evacuation of Black epistemic agency, demonstrates precisely this: both registers advocate for more accurate representations of Blackness while maintaining the structural condition in which Blackness remains the object of representation and the Human its subject.

Causal Interpretation Across Both Corpora

The most analytically significant divergence between the two corpora emerged in causal interpretation — how authors in each register identified the forces producing genAI-related harms to Black communities. Where the media corpus consistently attributed harm to historical and systemic forces, the empirical corpus revealed a pattern of structural evacuation: the active removal of causal explanation from

research accounts in ways that reproduce what Dancy and Saucier (2022) describe as the epistemic tendency of AI ethics discourse to “skip over addressing Blackness and anti-Blackness directly”.

In the media corpus, causal attribution was both pervasive and structurally oriented. [21] of [28] resources identified historical bias as a causal force, using language — “historically,” “pre-existing,” “ingrained,” “embedded” — to locate the roots of genAI harm in centuries-old racialized logics that have persisted into contemporary technological systems. [25] of [28] resources identified systemic bias as a causal force, attributing harm to current policies, organizational cultures, and exclusionary practices in AI development that actively reproduce anti-Black conditions. Notably, this structural causal framing appeared not only in explicitly critical outlets but also in economic and business media (e.g., Forbes, McKinsey) that have no particular commitment to structural critique. When sources concerned primarily with economic productivity and workforce planning nonetheless identified historical and systemic bias as the forces creating harm, this signals that the structural account of causality is available and legible across the discourse on genAI and Black communities, rather than a specialized theoretical claim requiring particular expertise to make.

The SLR corpus diverged sharply. [30] of [91] papers included no causal interpretation at all, treating AI systems instrumentally, as tools to be evaluated or improved, without implicating any forces as responsible for the harms or limitations they identified. Among papers that did include causal interpretation, the most common account attributed bias to training data: [55] of [91] papers identified embedded bias in datasets as the causal mechanism, with language such as “given that VLMs are primarily trained on web-scraped data, which contains human-created content reflecting existing social biases” (Lee et al. 2024). The issue is not that research papers fail to write like newspaper features. It is that the causal accounts they do provide systematically stop at the dataset, identifying bias as a property of the training data without reflecting on the social conditions that produced it. Dancy and Saucier anticipate this move precisely, arguing that AI ethics discussions reproduce the tendency to focus on “individuals and specific acts or bias towards anti-Black, racist outcomes” while the “individual focus obscures a structural cause that historically has had a longer lasting and more pervasive impact” (Dancy and Saucier 2022). In the SLR corpus, the dataset functions as the individual-level unit of analysis: bias is located in a specific artifact rather than in the structuring conditions that made that artifact possible. The causal account shapes the treatment account, and a truncated causal account produces a truncated treatment.

Treatment Recommendation: Technical Reform, Structural Protection, and the Shared Evacuation of Black Epistemic Agency

The treatment recommendations across both corpora follow directly from their causal accounts. Where causal interpretation was truncated — stopping at the dataset rather than implicating the structural conditions that produced those data

— treatment recommendations concentrated on technical reform. Where causal interpretation extended to structural and historical forces, treatment recommendations reached toward regulation and systemic intervention.

In the SLR corpus, technical reform dominated overwhelmingly. [78] of [91] papers recommended improvements to AI development as their primary treatment, oriented toward better datasets, revised annotation practices, improved benchmarks, and algorithmic adjustments to reduce bias. No empirical paper recommended regulatory intervention, community-led governance, or structural changes to the conditions under which AI development occurs. Lakkaraju et al.’s financial advisory study (Lakkaraju et al. 2023) illustrates how this pattern operates at the level of individual research: test scenarios designed to evaluate LLM performance with AAVE speakers reproduced exaggerated linguistic stereotypes — “*I be makin’ a purchase of \$1000 usin’ i’s credit card*” — in the very process of assessing bias. A study oriented toward identifying harm in AI systems reproduced that harm through its own methodology, precisely because its remedial horizon extended no further than the system being evaluated.

The media corpus recommended more expansive treatments. [13] of [28] resources called for regulatory intervention, addressing legislators, educators, and institutional policymakers as agents capable of governing genAI’s impact on Black communities. [9] of [28] resources recommended participatory strategies to prevent or mitigate harm, with economic media in particular urging employers to “bring a participatory approach to AI planning” (Keenan 2023). Reform of AI development itself was also common, with representation among creators framed as a direct intervention: “Diversifying the creators of AI has a direct impact on the data that the algorithms produce” (Graves et al. 2024). One resource concluded with a call to resist, urging readers to refuse Meta’s AI products, “Either way, we need to be ready to resist whatever internet decay Meta has in store for us next” (Attiah 2025), framing individual refusal as a meaningful form of protection.

Yet even the media corpus’s more structurally oriented recommendations reproduce a specific limitation. Across all treatments — *expand access, regulate, resist, reform* — Black communities are consistently positioned as recipients of protective action undertaken by others: legislators, educators, developers, employers. Even resistance is framed as individual refusal rather than collective agency in shaping what AI becomes. What is absent from both corpora — and the absence is shared across both registers despite their different orientations — is any treatment recommendation that positions Black communities as epistemic agents in determining what genAI should do, how it should be governed, and who should hold authority over its development. While we could not code for community involvement in media resources, we found that [14] SLR papers engaged the community in their research, but in each case, involvement was framed as consultation or input into processes controlled elsewhere — participatory in method but not in authority. The treatment recommendations in both corpora, however structurally ambitious, maintain Blackness as the

object upon which AI systems act, rather than as the epistemological subject capable of determining the conditions of that action.

Discussion

Our findings reveal that the misalignment between public discourse and scholarly research on genAI and Black communities is not merely a matter of emphasis or priority; it is structurally produced. Where public discourse consistently attributed genAI-related harm to historical and systemic forces and reached toward regulatory and structural interventions, the SLR corpus reveals a focus that stops at causal accounts of datasets and treatment recommendations primarily involving technical reform. Reading both corpora through Dancy and Saucier’s 2022 onto-epistemological framework surfaces what the numbers alone cannot: anti-Blackness operates simultaneously across both registers, generating not only divergence but a shared evacuation of Black epistemic agency that persists even in the more structurally oriented framings of public discourse. This convergence — representational harm as the dominant problem definition in both corpora, accompanied by the shared absence of Black communities as participatory agents, knowledge producers and design authorities — is the analytic center of what follows.

These findings carry important limitations that bear on their interpretation. Media discourse, shaped by journalistic news values and editorial priorities, does not straightforwardly represent the perspectives of Black communities themselves; our analysis characterizes public narratives about Black communities and genAI rather than directly capturing the views of Black users and community members. Extending frame analysis to social media discourse—where Black users articulate their own framings of genAI rather than being framed by journalistic and institutional intermediaries—is a natural next step, and one particularly aligned with our call to center Black communities as epistemic agents. Capturing actual discourse at public forums and summits that have emerged around the topic of equitable AI for Black futures may also serve this purpose. Our review of empirical literature is similarly bounded by computing venues and the recency of genAI research, capturing an early-stage landscape that continues to evolve rapidly. What both limitations share is an absence of Black community voice — not only in the corpora we analyzed, but as a constitutive condition of how both public discourse and scholarly research are currently produced. It is precisely this absence that the following section takes up directly.

Attending to Blackness in GenAI Efforts

The analysis of both media sources and empirical literature related to genAI and Black communities points to an epistemological enclosure of Blackness, one that maintains racialized boundaries (Dancy and Saucier 2022) that ultimately position Black people, communities, and practices as objects, deviant, and lacking. Yet Blackness should not be relegated to a site of harm; it is also a site of knowledge, agency, and what Warren (Warren 2018) calls ‘ontological

possibility' in the face of anti-Black terror. Black radical traditions offer a different cultural grounding altogether, one in which liberatory futures are not extensions of existing systems but imaginative departures from them (Robinson 2020; Kelley 2022). That both public discourse and empirical research systematically foreclose this dimension of Blackness — its generative, world-making capacity — is therefore not a minor oversight but a structural condition worth naming directly. Though analysis of public discourse indicated accounts better positioned to address the multi-level, systemic issues at the intersection of Blackness and genAI than scholarly research, both accounts point to a broader issue: the recognition of the human agency of Black people and communities as they engage with increasingly embedded genAI systems.

The maintained enclosure of Blackness in spaces of other (than human) should, in and of itself, be more directly addressed in both public discourse and scholarly research related to genAI, but even this misses key considerations. First, by failing to register relations between Black being and (foundational conceptions of) the human (Warren 2018; Wynter 2003; Wynter and McKittrick 2015), both public discourse and research leave the trouble with concepts such as *freedom* and *equality* largely unattended. Second, the sole focus of Blackness as an epistemological space of dereliction in relation to (a genre of) the human limits otherwise possibilities and is a symptom of a predominantly externalized gaze and non-participatory engagement. Here, O'Byrne's (2024) discussion of Blackness, nihilism, and relations to the universalist assumptions of the human (humanism) is helpful as they note work on "Black radical reconstructions" of the human to address (anti-)Blackness where Black life may transcend anti-Blackness with "infinite possibilities for thriving". That is, Black radical, liberatory traditions (e.g., see Robinson 2020; Kelley 2022) provide a different cultural grounding that makes possible a conceptual and action space that gives way to understanding Blackness, relations to the human, and otherwise possibilities due to a different cultural-experiential conceptual grounding (Lakoff and Johnson 2003) within Black communities. By discounting the behaviors and practices-the living-that occur within Black communities in the face of anti-Blackness, both genAI public discourse and genAI empirical research preclude liberatory futures.

Considering Blackness as a starting place for liberatory tech futures also creates questions about the role of race and other social constructs in utopian futures such as the current technological landscape that is largely ingrained with AI. This consideration of Blackness suggests holding space for existing joy, play, and freedom - the human condition of Blackness. Although the more recent focus of both public discourse and empirical research is to engage with the full human condition of Blackness, this requires no longer leaning towards extractive models that reduce certain humans to mere data (Network 2025). Such a reorientation raises the question of whether there is a future in which genAI can be liberatory, rather than another focal point of harm and bias. Liberatory instantiations of genAI require both the interrogation and perhaps conditional acceptance of these tools,

as well as the imagining of possibilities that cultivate both belonging and freedom through participatory and considerate approaches. Comparing emphasis in public discourse to trends in empirical research, we see a focus of research aiming to undo the latent biases of technology racialization, centering artifacts of Black oppression. Transcending harm mitigation offers technologies that empower and embrace communities that have historically been harmed by technology (Croskey, Benjamin, and Jacobs 2026; Croskey et al. 2025a). This is where the distinction between participation and epistemic grounding becomes critical. Our findings document that even the most community-engaged papers in the SLR corpus remained participatory in method but not in authority — Black communities were consulted but did not determine the terms of inquiry. A liberatory orientation toward genAI requires something more fundamental than expanded participation: it requires treating Black ways of knowing, making, and relating as constitutive of what AI is designed to do and for whom. This is what DISCO Network means by 'Black Indigeneity' (Network 2025) as a starting place, not a demographic consideration or a diversity add-on, but an ontological grounding that reorients the foundational questions of AI development: what counts as a problem worth solving, whose knowledge authorizes a solution, and what futures are imaginable from the outset. What would it mean to have Blackness not as a corrective to AI's defaults, but as the default itself?

Conclusion

We examined how public discourse and empirical scholarly research align and misalign on genAI's impact on Black communities. Our framing analysis of public discourse revealed an emphasis on historical and structural dimensions of harm as an important causal force, while analysis of the scholarly research landscape revealed either an absence of causal explanation or shallow, technical, infrastructure-oriented explanations, along with a focus on technical (bias-oriented) reform in this causal treatment. Even in this divergence of causal treatment, both corpora reveal a limitation in that Black individuals and communities are mainly positioned as without authoritative agency in treatment recommendations, precluding the potential interventions towards otherwise liberatory futures.

Though public discourse can have direct and diffuse effects on scholarly research, with the former potentially anticipating material harms and consequences, positive impacts are not guaranteed, and research can be epistemologically structured to exclude such impacts. In addition to our contributions in public discourse and empirical scholarly literature analysis, we argue for frame analysis as an AI ethics methodology suited to the anticipatory, structural questions that venues like AIES exist to address. By using this method to understand the terrain of public concern and research focus, we contribute to the growing effort to ensure that AI research addresses issues pertinent to the communities most likely negatively impacted by these systems, and highlight the continued need for structural interrogations.

Positionality Statement

Our research team consists of six researchers spanning graduate student, academic faculty, and industry research positions. Most members of our team are of African and African American descent, with identities that intersect with other ethnicities as well, and our lived experience with anti-Blackness informs, but does not exhaust, our analytic commitments. Our expertise spans critical theories in HCI and design, community-based participatory research, AI research centering anti-Blackness, and applied work on AI harms. Several of us also study how to involve underrepresented groups in AI development and evaluation. We approached this work with the prior conviction — grounded in Black studies scholarship and in our own research trajectories — that Blackness is not reducible to a demographic variable, and that anti-Blackness is a structural condition rather than an artifact of flawed datasets, an orientation that shaped both how we coded the corpora and what we were positioned to notice in them.

Our institutional positions also condition this work. Members of our team are situated within academic institutions and within an industry research organization whose parent company develops and deploys the class of generative AI systems we critique. We recognize that this proximity affords access and insight while also constraining what is easy to say and from where; we have tried to hold that tension explicitly rather than resolve it, and to let the analysis follow the evidence in both corpora regardless of which institutions it implicates. Finally, we note that our team's composition does not substitute for the community authority we argue is absent from both registers. This paper analyzes discourse about Black communities; it does not speak for them, and our critique of consultation-without-authority applies reflexively to the limits of our own method.

References

- Abbas, C.; Awad, M.; and Tajeddine, R. 2025. Can small-scale data poisoning exacerbate dialect-linked biases in large language models? *arXiv [cs.CL]*.
- Abdur-Rahman, N. 2025. Billionaire Robert F. Smith brings Generative AI curriculum to Morehouse and Spelman Colleges.
- AlMakinah, R.; Norcini-Pala, A.; Disney, L.; and Canbaz, M. A. 2025. Enhancing mental health support through human-AI collaboration: Toward secure and empathetic AI-enabled chatbots. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, 196–202. IEEE.
- Asare, J. G. 2023. Is Artificial Intelligence Anti-Black?
- Attiah, K. 2025. I talked to Meta's Black AI character. Here's what she told me.
- Banks, M. 2025. Stargate AI: Will Black Americans Be Left Behind Again?
- Basoah, J.; Chechelnitsky, D.; Long, T.; Reinecke, K.; Zerva, C.; Zhou, K.; Díaz, M.; and Sap, M. 2025a. Not like us, hunty: Measuring perceptions and behavioral effects of minoritized anthropomorphic cues in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 710–745. New York, NY, USA: ACM.
- Basoah, J.; Cunningham, J. L.; Adams, E.; Bose, A.; Jain, A.; Yadav, K.; Yang, Z.; Reinecke, K.; and Rosner, D. 2025b. Should AI mimic people? Understanding AI-supported writing technology among black users. *Proc. ACM Hum. Comput. Interact.*, 9(7): 1–51.
- Baumer, E. P.; Mimno, D.; Guha, S.; Quan, E.; and Gay, G. K. 2017. Civic seed: How to engage with civic issues through technology design. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2010–2021.
- Bautista, Y. J. P.; Theran, C.; Aló, R.; and Lima, V. 2023. Health disparities through generative AI models: a comparison study using a Domain specific large Language Model. In *Proceedings of the Future Technologies Conference*, 220–232. Springer.
- Beer, J. 2025. AI generators say there are no Black surfers. This group is out to change that.
- Benjamin, R. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.
- Berkeley, N. 2025. Poisoned for Progress: How Ai is quietly draining Black Communities.
- Bhardwaj, N.; Bhardwaj, A.; and Garg, L. 2025. Controlling bias in generative AI: Techniques for fair and equitable data generation in socially sensitive applications. In *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS)*, 1–8. IEEE.
- Bilalpur, M.; Hamm, M.; Lee, Y. J.; Norman, N.; McTigue, K. M.; and Wang, Y. 2025. Topic-aware Large Language Models for Summarizing the Lived Healthcare Experiences Described in Health Stories. *arXiv preprint arXiv:2510.24765*.
- Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022a. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.
- Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022b. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 173–184.
- Blackstock, O.; and Fatou, A. 2025. Engage the community in Generative AI for public health.
- Blay, Z. 2025. Digital Blackface is back in the form of Black Ai Influencers.
- Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5454–5476.
- Bonilla-Silva, E. 2015. The Structure of Racism in Color-Blind, "Post-Racial" America. *American Behavioral Scientist*, 59(11): 1358–1376.
- Bosco, C.; Shojaei, F.; Theisz, A. A.; Nguyen, V.; Song, H.; Han, R.; Torres, J. A. O.; Chheda, D.; Lin, J.; Peng, X.;

- et al. 2025. "I don't see anything specifically about Black-/African Americans." Testing an Alzheimer-specific generative AI tool tailored for African American/Black communities. *ACM Transactions on Computing for Healthcare*.
- Bray, K.; and Harrington, C. 2021. Speculative Blackness: Considering Afrofuturism in the Creation of Inclusive Speculative Design Probes.
- Brewer, R. N.; Harrington, C.; and Heldreth, C. 2023. Envisioning equitable speech technologies for black older adults. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 379–388.
- Brown, J. S.; Finney, M.; Korgaonkar, N.; McMillan, M.; and Perkins, C. 2023. The impact of generative AI on Black Communities.
- Cantini, R.; Orsino, A.; Ruggiero, M.; and Talia, D. 2025. Benchmarking adversarial robustness to bias elicitation in large language models: scalable automated assessment with LLM-as-a-judge. *Mach. Learn.*, 114(11).
- Chang, T.; and Wiens, J. 2024. Racial differences in laboratory testing as a potential mechanism for bias in AI: a matched cohort analysis in emergency department visits. *PLOS Global Public Health*.
- Chauhan, A.; Anand, T.; Jauhari, T.; Shah, A.; Singh, R.; Rajaram, A.; and Vanga, R. 2024. Identifying race and gender bias in stable diffusion ai image generation. In *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, 1–6. IEEE.
- 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*.
- Chow, A. R. 2025. Inside Memphis' battle against Elon Musk's xai data center.
- Coggins, W.; McKenzie, J.; Youm, S.; Mummaleti, P.; Gilbert, J.; Ragan, E.; and Dorr, B. J. 2025. That Ain't right: Assessing LLM performance on QA in African American and west African English dialects. In *Proceedings of the 9th Widening NLP Workshop*, 123–129. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Croskey, P.; Benjamin, R.; and Jacobs, J. 2026. The Augmented Undercommons: A Framework for Liberatory HCI Research within Ethically Compromised Institutions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.
- Croskey, P.; Offert, F.; Jacobs, J.; and Thaler, K. M. 2025a. Liberatory Collections and Ethical AI: Reimagining AI Development from Black Community Archives and Datasets. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 900–913. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Croskey, P.; Offert, F.; Jacobs, J.; and Thaler, K. M. 2025b. Liberatory Collections and Ethical AI: Reimagining AI Development from Black Community Archives and Datasets. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 900–913. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Cunningham, J.; Benabdallah, G.; Rosner, D.; and Taylor, A. 2023. On the Grounds of Solutionism: Ontologies of Blackness and HCI. *ACM Trans. Comput.-Hum. Interact.*, 30(2): 1–17.
- Dancy, C. L.; and Saucier, P. K. 2022. AI and Blackness: Towards moving beyond bias and representation. *IEEE Transactions on Technology and Society*, 3(1): 31–40.
- Deas, N.; Grieser, J.; Kleiner, S.; Patton, D.; Turcan, E.; and McKeown, K. 2023. Evaluation of African American language bias in natural language generation. *arXiv preprint arXiv:2305.14291*.
- Dev, T.; Sultana, S.; and Bosu, A. 2025. Beyond binary moderation: Identifying fine-grained sexist and misogynistic behavior on GitHub with large language models. In *2025 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 01–12. IEEE.
- Devinney, H.; Björklund, J.; and Björklund, H. 2024. We don't talk about that: case studies on intersectional analysis of social bias in large language models. In *Workshop on Gender Bias in Natural Language Processing (GeBNLP), Bangkok, Thailand, 16th August, 2024.*, 33–44.
- Díaz, M.; Dev, S.; Reif, E.; Denton, E.; and Prabhakaran, V. 2024. SoUnD Framework: Analyzing (So) cial Representation in (Un) structured (D) ata. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 371–383.
- Dildora, I.; Shokirovich, O. K.; Yuldasheva, G.; Singhal, A.; Alimardanova, R.; and Abdirahmonov, L. 2025. Democratizing AI Linguistic Technologies for Understanding Diverse Social Voices. In *2025 7th International Conference on Information Systems and Computer Networks (ISCON)*, 1–5. IEEE.
- Dinkar, T.; Jiang, A.; Abercrombie, G.; and Konstas, I. 2025. Erasing 'ugly' from the internet: Propagation of the beauty myth in text-image models. *arXiv [cs.CV]*.
- Dorn, R.; Chance, C.; Rusti, C.; Bickham Jr, C.; Chang, K.-W.; Morstatter, F.; and Lerman, K. 2025. Reinforcing Stereotypes of Anger: Emotion AI on African American Vernacular English. *arXiv preprint arXiv:2511.10846*.
- Dossou, B. F.; and Aidasso, H. 2025. Towards Open-Ended Discovery for Low-Resource NLP. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, 287–297.
- Eledroos, N. 2025. Progress shouldn't Poison Black Communities.
- Elsharif, W.; Alzubaidi, M.; and Agus, M. 2025. Cultural bias in text-to-image models: A systematic review of bias identification, evaluation, and mitigation strategies. *IEEE Access*, 13: 122636–122659.
- Enterprise, B. 2025. How Amazon's Generative AI is empowering black innovators entrepreneurs.
- Entman, R. M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4): 51–58.

- Eschner, J.; Labadie-Tamayo, R.; Zeppelzauer, M.; and Waldner, M. 2025. Interactive discovery and exploration of visual bias in generative Text-to-Image models. *Comput. Graph. Forum*, 44(3).
- Finch, S. E.; Paek, E. S.; Choi, I.; and Choi, J. D. 2025. Finding A voice: Exploring the potential of African American dialect and voice generation for chatbots. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 25789–25806. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Fu, Z.; Brown, R.; Shao, S.; Rawal, K.; Delaney, E.; and Russell, C. 2025. FairImagen: Post-processing for bias mitigation in text-to-image models. *arXiv [cs.LG]*.
- Gaskins, N. 2024. Trilogy of Difference: Black Art, Femininity & Generative AI.
- Graves, E.; Aswar, S.; Desai, R.; Nampelli, S.; Chakraborty, S.; and Hall, T. 2024. Aave corpus generation and low-resource dialect machine translation. In *Proceedings of the 7th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, 50–59.
- Greene-Santos, A. 2024. Does AI Have a Bias Problem?
- Groenwold, S.; Ou, L.; Parekh, A.; Honnavalli, S.; Levy, S.; Mirza, D.; and Wang, W. Y. 2020. Investigating African-American Vernacular English in transformer-based text generation. *arXiv preprint arXiv:2010.02510*.
- Guo, W.; and Caliskan, A. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 122–133.
- Gupta, A.; Meng, P.; Yurtseven, E.; O'Brien, S.; and Zhu, K. 2024. AAVENUE: Detecting LLM Biases on NLU Tasks in AAVE via a Novel Benchmark. *arXiv:2408.14845*.
- Gupta, S. 2024. AI generates harsher punishments for people who use Black dialect.
- Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.
- Gusenbauer, M.; and Haddaway, N. R. 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research synthesis methods*, 11(2): 181–217.
- Haider, B.; Gorti, A.; Chadha, A.; and Gaur, M. 2025. Mental health equity in LLMs: Leveraging multi-hop question answering to detect amplified and silenced perspectives. *arXiv [cs.CL]*.
- Haimson, O. L.; Mayworm, S. R.; Ingber, A. S.; and Andalibi, N. 2025. AI Attitudes Among Marginalized Populations in the U.S.: Nonbinary, Transgender, and Disabled Individuals Report More Negative AI Attitudes. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 1224–1237. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Hale, K. 2024. Generative AI's \$40 Billion Obstacle For Black Workers.
- Hanna, A.; Denton, R.; Smart, A.; and Smith-Loud, J. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 501–512.
- Harris, C.; Johnson, A. G.; Palmer, S.; Yang, D.; and Bruckman, A. 2023. "Honestly, I Think TikTok has a Vendetta Against Black Creators": Understanding Black Content Creator Experiences on TikTok. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–31.
- Harvey, E.; Kizilcec, R. F.; and Koenecke, A. 2025. A framework for auditing chatbots for dialect-based quality-of-service harms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025–2039. New York, NY, USA: ACM.
- Hassan, M. F.; Khattak, F. K.; and Seyyed-Kalantari, L. 2025. Dialectic Preference Bias in Large Language Models. In *Proceedings of the AAAI Symposium Series*, volume 5, 365–369.
- He, T.-T.; Malandro, M. E.; and Shadle, D. 2025. The Florence Price Art Song Dataset and Piano Accompaniment Generator. *arXiv preprint arXiv:2506.23130*.
- Hoehn, S.; Philipppy, F.; and Andre, E. 2025. On Speakers' Identities, Autism Self-Disclosures and LLM-Powered Robots. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 484–503.
- Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028): 147–154.
- Howard, P.; Madasu, A.; Le, T.; Moreno, G. L.; Bhiwandwalla, A.; and Lal, V. 2024. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11975–11985.
- Huang, H.; Jin, X.; Miao, J.; and Wu, Y. 2025. Implicit Bias Injection Attacks against Text-to-Image Diffusion Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28779–28789.
- Jung, S.; Oesterling, A.; Verdun, C. M.; Vithana, S.; Moon, T.; and Calmon, F. P. 2025. Multi-Group Proportional Representation for text-to-image models. *arXiv [cs.CV]*.
- Junias, O.; Kini, P.; and Chaspari, T. 2025. Assessing algorithmic bias in language-based depression detection: A comparison of DNN and LLM approaches. *arXiv [cs.CL]*.
- Kamruzzaman, M.; and Kim, G. L. 2025. Evaluating the impact of racial cues on MLLMs judgements of politeness and offensiveness. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 7713–7722. IEEE.
- Kay, J.; Kasirzadeh, A.; and Mohamed, S. 2024. Epistemic injustice in generative AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 684–697.

- Keenan, S. 2023. Report Shows: 53% Of Black Workers Fear AI Job Displacement, Yet 45% Embrace AI Tools At Work.
- Kelley, P. G.; Yang, Y.; Heldreth, C.; Moessner, C.; Sedley, A.; Kramm, A.; Newman, D. T.; and Woodruff, A. 2021. Exciting, useful, worrying, futuristic: Public perception of artificial intelligence in 8 countries. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 627–637.
- Kelley, R. D. G. 2022. *Freedom dreams: The Black radical imagination*. Boston, MA, USA: Beacon Press.
- Kempermann, M.; Vasu, S. S. M.; Raveenthiran, M.; Farrell, T.; and Weber, I. 2026. Safe for Whom? Rethinking How We Evaluate the Safety of LLMs for Real Users. *arXiv:2512.10687*.
- Kendi, I. X. 2017. *Stamped from the beginning: The definitive history of racist ideas in America*. New York, NY: Bold Type Books. ISBN 1473549477.
- Khorramrouz, A.; and Levy, S. 2025. Characterizing selective refusal bias in large language models. *arXiv [cs.CL]*.
- Kim, S.; Wang, J.; Janssen, M. A.; and Anderies, J. M. 2025. How large language models systematically misrepresent American climate opinions. *arXiv [cs.CY]*.
- Klein, A. 2024. Black Students Are More Likely to Be Falsely Accused of Using AI to Cheat.
- Kondrup, E.; and Imouza, A. 2025. Dr. Bias: Social disparities in AI-powered medical guidance. *arXiv [cs.AI]*.
- Kuck, K. 2023. Generative artificial intelligence: A double-edged sword. In *2023 World Engineering Education Forum-Global Engineering Deans Council (WEEF-GEDC)*, 1–10. IEEE.
- Lakkaraju, K.; Vuruma, S. K. R.; Pallagani, V.; Muppasani, B.; and Srivastava, B. 2023. Can llms be good financial advisors?: An initial study in personal decision making for optimized outcomes. *arXiv preprint arXiv:2307.07422*.
- Lakoff, G.; and Johnson, M. 2003. *Metaphors We Live By*. Chicago, IL, USA: The University of Chicago Press.
- Leavy, S.; Siapera, E.; and O’Sullivan, B. 2021. Ethical data curation for AI: An approach based on feminist epistemology and critical theories of race. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 695–703.
- Lee, J.; Shang, T.; Baik, J. Y.; Duong-Tran, D.; Yang, S.; Li, L.; and Shen, L. 2025. From promising capability to pervasive bias: Assessing large Language Models for emergency department triage. *arXiv [cs.AI]*.
- Lee, M. H.; and Jeon, S. 2024. Vision-Language Models Represent Darker-Skinned Black Individuals as More Homogeneous than Lighter-Skinned Black Individuals. *arXiv preprint arXiv, 2412*.
- Lee, M. K.; and Rich, K. 2021. Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–14.
- Lee, S.; Peng, T.-Q.; Goldberg, M. H.; Rosenthal, S. A.; Kotcher, J. E.; Maibach, E. W.; and Leiserowitz, A. 2024. Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. *PLoS Climate*, 3(8): e0000429.
- Liang, Z.; Suresh, A.; and Chen, I. Y. 2025. Revealing treatment non-adherence bias in clinical machine learning using large language models. *arXiv [cs.LG]*.
- Lima, G.; Grgić-Hlača, N.; and Cha, M. 2023. Blaming humans and machines: What shapes people’s reactions to algorithmic harm. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–26.
- Liu, Z.; Li, H.; Chen, A.; Zhang, R.; and Lee, Y.-C. 2024. Understanding public perceptions of AI conversational agents: A cross-cultural analysis. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, 1–17.
- Lutz, M.; Sen, I.; Ahnert, G.; Rogers, E.; and Strohmaier, M. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 23212–23237. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Malik, A.; Sabri, N.; Karnaze, M. M.; and ElSherief, M. 2025. Are LLMs empathetic to all? Investigating the influence of multi-demographic personas on a model’s empathy. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 24938–24959. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Maluleke, V. H.; Thakkar, N.; Brooks, T.; Weber, E.; Darrell, T.; Efros, A. A.; Kanazawa, A.; and Guillory, D. 2022. Studying bias in gans through the lens of race. In *European Conference on Computer Vision*, 344–360. Springer.
- Marrinan, C. 2025. Data Center Boom Risks Health of AI-ready Vulnerable Communities.
- Matthes, J.; and Kohring, M. 2008. The content analysis of media frames: Toward improving reliability and validity. *Journal of communication*, 58(2): 258–279.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.
- Mire, J.; Aysola, Z. T.; Chechelnitsky, D.; Deas, N.; Zerva, C.; and Sap, M. 2025. Rejected dialects: Biases against african american language in reward models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 7468–7487.
- Morales, S.; Clarisó, R.; and Cabot, J. 2025. ImageBiTe: A framework for evaluating representational harms in text-to-image models. In *2025 IEEE/ACM 4th International Conference on AI Engineering-Software Engineering for AI (CAIN)*, 95–106. IEEE.
- Morrow, A. 2025. Meta scrambles to delete its own AI accounts after backlash intensifies — CNN business.
- Narayan, M.; Pasmore, J.; Sampaio, E.; Raghavan, V.; Maity, S.; Waters, G.; and Howard, A. M. 2025. Mitigating bias in large language models through culturally-relevant LLMs. In *2025 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)*, 1–7. IEEE.

- Neely, A.; and Yorke, T. 2025. Generative AI and virtual reality for scientific literacy and STEM identity development. In *2025 Black Issues in Computing Education (BICE)*, 16–20. IEEE.
- Network, D. 2025. *Technoskepticism: Between possibility and refusal*. Stanford University Press.
- Nezhad, S. B.; Bandyapadhyay, S.; and Agrawal, A. 2024. Fair Summarization: Bridging Quality and Diversity in Extractive Summaries. *arXiv preprint arXiv:2411.07521*.
- Nias, J.; Brandon, D.; Douglass, T.; Haynes, A.; Patterson, N.; and Cooper, K. 2025. From culture to code: An inter-sectional analysis of AAVE slang in large language models. In *Proceedings of the Fifth Biennial African Human-Computer Interaction Conference*, 141–150. New York, NY, USA: ACM.
- Noble, S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
- Ocuppaugh, J.; Liu, X.; and Zambrano, A. F. 2025. Language models and dialect differences. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 204–215. New York, NY, USA: ACM.
- Ogbonnaya-Ogburu, I. F.; Smith, A. D. R.; To, A.; and Toyama, K. 2020. Critical race theory for HCI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–16.
- Okpala, E.; and Cheng, L. 2025. Large language model annotation bias in hate speech detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 19: 1389–1418.
- O’Leary, T. K.; Batti, V.; Dhindsa, S.; Krishnan, V. H.; Wilcox, L.; Woods-Giscombe, C. L.; Bellamy, C. D.; and McCall, T. 2025. “It was Empowering”: Care Pathways a Novel Culturally Informed Inspiration Card Game for Community-Engaged Co-Design with Black American Women and Clinicians. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA ’25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.
- Olojo, S.; Zakrzewski, J.; Smart, A.; van Lient, E.; Miceli, M.; Ebinama, A.; and Amugongo, L. M. 2025. Lost in Machine Translation: The Sociocultural Implications of Language Technologies in Nigeria. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2249–2259.
- Osti, G.; and Roke, E. R. 2024. Collaborating for Change? Assessing Metadata Inclusivity in Digital Collections with Large Language Models (LLMs). In *2024 IEEE International Conference on Big Data (BigData)*, 2479–2488.
- O’Byrn, E. 2024. Nothing and Infinity: Black Life’s Response to Ontological Terror. *Critical Philosophy of Race*, 12(2): 382–400.
- Parker, A. G.; Vardoulakis, L. M.; and Harrington, C. N. 2026. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI ’26’. New York, NY, USA: Association for Computing Machinery.
- Pham, H.; Kohli, T.; Llano, E. O.; Nokuri, I.; and Weinstock, A. 2024. How will AI Impact Racial Disparities in Education?
- Porwal, R.; Rozet, A.; Gowda, J.; Houck, P.; Tang, K.; and Moeller, S. 2025. Analysis of LLM as a grammatical feature tagger for African American English. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 7744–7756. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Proebsting, G.; Anigboro, O. I.; Crawford, C. M.; Metaxa, D.; and Friedler, S. A. 2025. Identity-related speech suppression in generative AI content moderation. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 185–217. New York, NY, USA: ACM.
- Qadri, R.; Diaz, M.; Wang, D.; and Madaio, M. 2025. The case for “thick evaluations” of cultural representation in ai. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.
- Rastogi, C.; Teh, T. H.; Mishra, P.; Patel, R.; Wang, D.; Díaz, M.; Parrish, A.; Davani, A. M.; Ashwood, Z.; Paganini, M.; Prabhakaran, V.; Rieser, V.; and Aroyo, L. 2025. Whose view of safety? A deep DIVE dataset for pluralistic alignment of text-to-image models. *arXiv [cs.LG]*.
- Ray, V. 2019. A Theory of Racialized Organizations. *American Sociological Review*, 84(1): 26–53.
- Ren, R.; Yao, X.; Cole, S.; and Wang, H. 2024. Are Large Language Models Ready for Travel Planning? *arXiv preprint arXiv:2410.17333*.
- Richards, W. b. A. 2025. Racist AI-generated videos are the newest slop garnering millions of views on TikTok.
- Robinson, C. J. 2020. *Black marxism, revised and updated third edition: The making of the black radical tradition*. Chapel Hill, NC, USA: The UNC Press, 3 edition.
- Saffari, H.; Shafiei, M.; Zhang, H.; Harris, L.; and Moosavi, N. S. 2024. Beyond hate speech: NLP’s challenges and opportunities in uncovering dehumanizing language. *arXiv [cs.CL]*.
- Sakunkoo, A.; and Sakunkoo, J. 2025. Name of thrones: How do LLMs rank student names in status hierarchies based on race and gender? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, 697–707. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Sandoval, S. C.; Acquaye, C.; Cobbina, K. A.; Teli, M. N.; and Daumé III, H. 2025. My LLM might Mimic AAE-But When Should it? 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)*, 5277–5302.
- Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

- Scheuerman, M. K.; Hanna, A.; and Denton, R. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–37.
- Septiandri, A. A.; Constantinides, M.; and Quercia, D. 2024. The Impact of Responsible AI Research on Innovation and Development. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1329–1342.
- Shankar, N. B.; Johnson, A.; Chance, C.; Veeramani, H.; and Alwan, A. 2024. CORAAL QA: A dataset and framework for open domain spontaneous speech question answering from long audio files. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 13371–13375. IEEE.
- Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Ros-tamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.
- Shukla, P. 2025. Understanding and evaluating computer vision models through the lens of counterfactuals. *arXiv [cs.CV]*.
- Shukla, P.; Chinchure, A.; Diana, E.; Tolbert, A.; Hosana-gar, K.; Balasubramanian, V. N.; Sigal, L.; and Turk, M. A. 2025. Mitigate one, skew another? Tackling intersectional biases in text-to-image models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 12453–12472. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Simone Noveck, B. 2025. Research radar: Race, democracy, and Ai - Spencer Overton offers a framework for a more inclusive Digital Future.
- Simpson, A. 2024. What will generative AI mean for the racial wealth gap?
- Slaughter, I.; Greenberg, C.; Schwartz, R.; and Caliskan, A. 2023. Pre-trained speech processing models contain human-like biases that propagate to speech emotion recognition. *arXiv preprint arXiv:2310.18877*.
- Sloane, M.; Moss, E.; Awomolo, O.; and Forlano, L. 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*, 1–6.
- Small, Z. 2023. Black Artists Say A.I. Shows Bias, With Algorithms Erasing Their History.
- Smith, A. D. R.; Ahmed, A. A.; Alvarado Garcia, A.; Dosono, B.; Ogbonnaya-Ogburu, I.; Rankin, Y.; To, A.; and Toyama, K. 2020. What’s race got to do with it? Engaging in race in HCI. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 1–8.
- Solyst, J.; Fong, C.; Nurdin, F.; Landesman, R.; and Benjamin Shapiro, R. 2025. “Grillz on a hijabi”: Intersectional Identities in Fostering Critical AI Literacy. *arXiv [cs.HC]*.
- Srinivasan, M.; and Abdel, J. 2025. GenFair: Systematic test generation for fairness fault detection in large language models. In *2025 IEEE International Conference on Artificial Intelligence Testing (AITest)*, 182–190. IEEE.
- Stefflbauer, N. 2024. Whose Job Is It To Protect Black People From AI? — Opinion.
- Stewart III, S. 2024. Gen Ai is Rising fast: How we ensure Black Americans rise with it.
- Suresh, H.; and Gutttag, J. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9.
- Swati, S.; Roy, A.; Panagiotou, E.; and Ntoutsis, E. 2025. MMM-fair: An interactive toolkit for exploring and operationalizing multi-fairness trade-offs. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 6713–6717. New York, NY, USA: ACM.
- Tanksley, T.; Smith, A. D.; Sharma, S.; and Huff Jr, E. W. 2025. “Ethics is not neutral”: Understanding Ethical and Responsible AI Design from the Lenses of Black Youth. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–20.
- Van Hulst, M.; Metze, T.; Dewulf, A.; De Vries, J.; Van Bommel, S.; and Van Ostaijen, M. 2025. Discourse, framing and narrative: three ways of doing critical, interpretive policy analysis. *Critical policy studies*, 19(1): 74–96.
- Verma, P.; Goyal, P.; Sharma, R.; Rani, D.; Raghuvanshi, R.; and Gupta, G. 2025. Language processing in AI and its impact on human identity and culture. In *2025 5th International Conference on Intelligent Technologies (CONIT)*, 1–7. IEEE.
- Verma, P.; and Unnikrishnan, B. 2025. When AI democratizes exploitation: LLM-assisted strategic manipulation of fair division algorithms. *arXiv [cs.CY]*.
- Wan, Y.; and Chang, K.-W. 2025. White men lead, black women help? Benchmarking and mitigating language agency social biases in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9082–9108.
- Wan, Y.; Subramonian, A.; Ovalle, A.; Lin, Z.; Suvarna, A.; Chance, C.; Bansal, H.; Pattichis, R.; and Chang, K.-W. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*.
- Wang, A.; and Zhai, L. 2024. GA-Net: Global-aware Attention-guided CNN for Food Image Classification. In *2024 IEEE World AI IoT Congress (AIIoT)*, 0408–0413. IEEE.
- Wang, X.; Xie, H.; Wang, Y.; Xiao, C.; Chen, H.; Sargeant, H.; Steffek, F.; Shao, J.; Liu, Z.; and Sun, M. 2025. Large language models’ complicit responses to illicit instructions across Socio-legal contexts. *arXiv [cs.CY]*.
- Warren, C. L. 2018. *Ontological Terror: Blackness, Nihilism, and Emancipation*. Duke University Press. ISBN 978-0-8223-7184-7.
- Webster, K. T. 2025. Fairness is not enough: Auditing competence and intersectional bias in AI-powered resume screening. *arXiv [cs.CY]*.

- Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.
- Williams, S. E. 2024. AI Revolution Threatens Job Security and Equity for Black Workers.
- Wilson, K.; Ghosh, S.; and Caliskan, A. 2025a. Bias amplification in Stable Diffusion’s representation of stigma through skin tones and their homogeneity. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2705–2717.
- Wilson, K.; Ghosh, S.; and Caliskan, A. 2025b. Bias Amplification in Stable Diffusion’s Representation of Stigma Through Skin Tones and Their Homogeneity. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 2705–2717.
- Wilson, K.; Sim, M.; Gueorguieva, A.-M.; and Caliskan, A. 2025. No thoughts just AI: Biased LLM hiring recommendations alter human decision making and limit human autonomy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2692–2704.
- Workman, D.; and Dancy, C. L. 2026. Identifying Potential Inlets of the Biocentric Man in the Artificial Intelligence Development Process. In Yu, P. K.; and Lim, D., eds., *Inclusive Innovation in the Age of AI and Big Data*, book section 10, 211–230. Oxford, UK: OUP. ISBN 9780197799413.
- Wu, Z.; Bulathwela, S.; Perez-Ortiz, M.; and Koshiyama, A. S. 2024. Stereotype Detection in LLMs: A Multi-class, Explainable, and Benchmark-Driven Approach. *arXiv preprint arXiv:2404.01768*.
- Wynter, S. 2003. Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation - An Argument. *CR: The New Centennial Review*, 3(3): 257–337.
- Wynter, S.; and McKittrick, K. 2015. *Unparalleled Catastrophe for Our Species? Or, to Give Humanness a Different Future: Conversations*, book section 2, 9–89. Durham, NC, USA: Duke University Press. ISBN 978-0-8223-5820-6.
- Xu, C. 2023. HarmonyVisage: Ethical Facial Emotion Dataset Using Advanced Generative AI. In *2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 0558–0563. IEEE.
- Xu, W.; Yan, Y.; and Zhu, Y. 2025. Bias in decision-making for AI’s ethical dilemmas: A comparative study of ChatGPT and Claude. *arXiv [cs.CY]*.
- Yanaka, H.; He, X.; Jie, L.; Han, N.; Oh, S.; Kumon, R.; Matsuoaka, Y.; Watabe, K.; and Itatsu, Y. 2025. Intersectional bias in Japanese large language models from a contextualized perspective. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 18–32. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Yucer, S.; Akçay, S.; Al-Moubayed, N.; and Breckon, T. P. 2020. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 18–19.
- Zade, H.; Williams, S.; Tran, T. T.; Smith, C.; Venkatagiri, S.; Hsieh, G.; and Starbird, K. 2024. To Reply or to Quote: Comparing Conversational Framing Strategies on Twitter. *ACM J. Comput. Sustain. Soc.*, 2(1).
- Zayed, A.; Mordido, G.; Shabanian, S.; Baldini, I.; and Chandar, S. 2024. Fairness-aware structured pruning in transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22484–22492.
- Zhang, A.; Yuksekogonul, M.; Guild, J.; Zou, J.; and Wu, J. C. 2023. ChatGPT exhibits gender and racial biases in acute coronary syndrome management. *arXiv preprint arXiv:2311.14703*.
- Zhang, Y.; Nguyen-Le, P.-A.; Singh, K.; and Gao, G. 2025. The News Says, the Bot Says: How Immigrants and Locals Differ in Chatbot-Facilitated News Reading. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.
- Zhao, Z.; Liu, Z.; Cao, Y.; Gong, S.; and Patras, I. 2025. AIM-Fair: Advancing algorithmic fairness via selectively fine-tuning biased models with contextual synthetic data. *arXiv [cs.CV]*.
- Zhou, M.; Abhishek, V.; Derdenger, T.; Kim, J.; and Srinivasan, K. 2024. Bias in generative AI. *arXiv preprint arXiv:2403.02726*.
- Zhou, R.; Wan, G.; Gabriel, S.; Li, S.; Gates, A. J.; Sap, M.; and Hartvigsen, T. 2025. Disparities in LLM reasoning accuracy and explanations: A case study on African American English. *arXiv [cs.CL]*.