Re-imagining Virtual Communities: Ethical Guidelines for Studying Black Twitter

Christina Chance, Kai-Wei Chang

University of California, Los Angeles
{cchance, kwchang}@cs.ucla.edu

Abstract

Black Twitter is an informal online network of Black users who leverage Twitter to share perspectives, build community, and mobilize around cultural and social justice issues. Black Twitter is a unique socio-cultural space where collective identity, shared experience, and cultural production converge. In this empirical study, we discuss Black Twitter as a virtual community, an online field site shaped by social interaction and platform dynamics, drawing on digital ethnographic methods like participant observation to understand its dynamics and the socio-technical systems that shape it. While existing frameworks and checklists fall short in providing methods for specifically studying virtual communities such as ignoring concerns of extracting data from closed communities and pushing for open access, we introduce an ethics-centered checklist for studying Black Twitter and more generally, marginalized virtual communities by addressing risks such as data misuse, data ownership, and misrepresentation. Applying this checklist, we conduct case studies to: 1) analyze how automatic moderation systems impact Black Twitter users' experiences and 2) explore in-group agreement on ownership and usage of reclaimed language. We find that moderation systems often misinterpret culturally-specific language and norms around reclaimed terms. To illustrate how users adapt language to circumvent flawed moderation systems, we perform keyword analysis to reveal that character-level perturbations of the communities' reclaimed slur reduces toxicity scores by 30.7%. Additionally, we conduct a community-sourced survey in which responses show that views on reclaimed slurs vary, with some linking them to the African diaspora and others to Black American identity, underscoring the need for culturally-aware moderation. Ultimately, our checklist offers an actionable framework for researchers to ethically engage with marginalized virtual communities emphasizing cultural nuance, accountability, and self-awareness.

Introduction

Black Twitter represents a vibrant virtual community where collective identity, shared experiences, and cultural production converge to create a socio-cultural hub for members of the African Diaspora. It is a space that highlights both the potential of marginalized communities to build solidarity and the challenges posed by the algorithms that surveil, shape, and sometimes suppress their voices. In a 2019 interview with Baylor University Media and Public Relations, Dr. Mia Moody-Ramirez, co-author of *From Blackface to Black Twitter: Reflections on Black Humor, Race, Politics, & Gender* and professor of Journalism, Public Relations and New Media at Baylor University, described Black Twitter as

> a grassroots movement within Twitter that has provided a virtual community of mostly African-American Twitter users a collective voice on a variety of issues, including Black Lives Matter. Black Twitter users often identify themselves using the #blacktwitter hashtag or by focusing on issues related to the black experience (White 2019).

This articulation underscores the importance of Black Twitter not only as a cultural and political force, but also, from a research standpoint, as a rich and influential online community that drives internet culture, supports political mobilization, and offers a valuable site for cross-disciplinary scholarly inquiry. Existing at the intersection of media studies, race and ethnicity studies, linguistics, and computer science, this community has the potential to inform and bring insight to many existing questions that these disciplines pose (i.e. What role do marginalized virtual communities play in agenda-setting and framing during political and social movements? How are distinctive features of African American Vernacular English (AAVE) introduced on Twitter and influence internet language?). While Black Twitter has been examined meaningfully across these disciplines, with many scholars acknowledging the ethical complexities of studying a marginalized online community (Mott and and 2021; Elwood and and 2018; Hair and Clark 2007; McInroy 2016), the field of AI has largely fallen behind in adopting similar practices and approaches (Lee, Jung, and Oh 2023; Dev et al. 2021; Koenecke et al. 2020).

Within AI and machine learning (ML), Black Twitter and other marginalized digital communities are often treated as mere data sources rather than as communities, with little regard for the potential harms of extraction, misrepresentation, generalization, or misuse. These harms range from the appropriation of entire online personas to seed synthetic data for generative models, to LLMs generating text about closed religious practices in Indigenous languages, sometimes exposing cultural secrets (Farahani and Ghasemi

2024; Haj Ahmad et al. 2025; Arora et al. 2023; Colón Vargas 2024; Olson, Guzmán, and Kunneman 2023). The ML community has built socio-technical infrastructures on the backs of data extracted from marginalized communities, often without adequate ethical precautions or care to protect them (McElroy and Vergerio 2022; Farayola et al. 2023; Angwin et al. 2016). As a result, these communities have grown wary and distrustful of AI and ML systems (Latham and Crockett 2024). They are often forced to find ways to circumvent or resist these technologies, or to bear the burden of discovering and advocating for solutions to the harms and biases these systems perpetuate (Knowles et al. 2023; Katyal 2022). This lack of ethical engagement not only risks reinforcing systemic biases but also overlooks the cultural, political, and emotional labor used to build and maintain these spaces. As a result, AI often fails to account for the impact that model development, deployment, and evaluation can have on the very communities it draws from.

Several checklists, frameworks, and guidelines have been proposed to support ethical, community-aware AI research. Many emphasize the importance of engaging with studied communities through participatory methods across the research pipeline (Olson, Guzmán, and Kunneman 2023; Parthasarathy and Katzman 2024; Ungless et al. 2025). These works often highlight the need for community feedback on research questions and methods, as well as the formation of community partnerships to validate the reliability of decisions made throughout the process. Other research focuses on data stewardship, advocating for responsible data collection practices such as respecting copyright and terms of use, accurately documenting the population from which the data originates, and being mindful of potential misuse or exposure of the studied community (Rogers, Baldwin, and Leins 2021). Some scholars have even challenged dominant notions of autonomy and self-determination, arguing that these ideals can further marginalize underrepresented groups. Instead, they call for a shift toward relational autonomy as part of broader efforts to decolonize AI (Mhlambi and Tiribelli 2023).

While these and other ethical AI guidelines exist, they are often too broad to address the specific challenges of researching marginalized digital communities. In this paper, we propose a community-informed ethics checklist tailored to studying Black Twitter as a virtual community. Our goal is to offer actionable guidance for responsible engagement, data stewardship, and navigating the ethical complexities of virtual ethnography, particularly in contexts where algorithmic systems may amplify harm. Through this work we explore the following questions: How can Black Twitter be conceptualized as a virtual community? What ethical challenges arise when studying marginalized online communities? How do algorithms affect the visibility and interactions within Black Twitter?

Additionally, we will situate Black Twitter within broader discussions of content moderation and platform governance, advocating for community-centered approaches that amplify marginalized voices. We utilize this checklist to conduct a small mixed-method study on the usage of the n-word within Black Twitter. Through this case study we perform a qualitative analysis of survey responses centered around the topic of reclaimed slur ownership and in-group vs out-group usage of the n-word. We found differing views on ownership of the reclaimed slur, with many respondents sharing that the historical context of this word influence what specific group this word is tied to. These responses aligned with general acceptability and usage of the word, noting that 80% of respondents stated that individuals out-of-group should never use the reclaimed slur. Additionally, we analyzed about 1.7 thousand tweets that utilized the n-word, assessing the sensitivity of Detoxify, a popular toxicity classifier, around the use of linguistic self-censorship, the misspelling and perturbation of words as a method to evade content moderation, as a means to circumnavigate repressive and bias content moderation systems. We found an alignment with the use of self-censorship and lower identity attack scores compared to the uncensored use of the word. Specifically, when we augmented the tweet from an uncensored use to an censored use, the self-censorship reduced the identity attack score on average by 30.7%. This suggested 1) that Detoxify heavily relies on reclaimed slur usage as a feature for classification and 2) that the use of self-censorship seems to successfully reduce the likelihood of a tweet being classified as an identity attack.

This research contributes to the larger body of work on virtual communities by framing Black Twitter as a case study that highlights the need for socio-technical systems that prioritize equity and inclusion. The long term goal is to explore how platform design can be improved to better serve marginalized groups and address the ethical and technological challenges they face.

Background

Concept of a Virtual Community The concept of virtual communities has been explored in numerous scholarly works, offering frameworks for understanding how digital spaces facilitate social connections and shared identities. Rheingold (2000) defines virtual communities as "social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace." This definition emphasizes the importance of emotional bonds and information exchange as core features of virtual communities. Ellis, Oldridge, and Vasconcelos (2004) builds on this, noting that virtual communities also balance the provision of social support with the sharing of personal experiences, which helps to create a sense of belonging. Kuntsman (1970), in her study of a virtual community of Queer Russian GLBT individuals in Israel, highlights how societal hierarchies and unspoken community rules persist even in digital spaces. These foundational works provide the tools to analyze the features necessary for defining virtual communities.

In the case of Black Twitter, it is a space where members can share personal experiences, engage in political discourse, and perform collective identity. These activities are central to understanding Black Twitter as a virtual community. By leveraging language, humor, and cultural practices,

members navigate a space that both shapes and is shaped by their collective experiences as part of the African Diaspora.

Algorithms and Marginalized Communities Within the virtual space, algorithms play an increasingly important role in shaping these communities often reinforcing and amplifying biases that disproportionately affect marginalized groups. Noble (2018) discusses how algorithms encode systemic inequities, particularly within search engines and social media platforms. She argues that these algorithms prioritize dominant narratives while suppressing marginalized voices, which exacerbates issues of visibility through systemic biases. Bender et al. (2021) extend this argument to large-scale language models, showing how AI systems trained on biased datasets further perpetuate these inequities. The result is an amplification of societal biases and a distortion of the representation of marginalized groups, including Black Twitter users.

In the context of Black Twitter, algorithms can significantly influence which voices are heard, which trends gain traction, and how discourse is shaped. The amplification of certain narratives—especially those that align with mainstream cultural norms—can suppress the complex, diverse, and often resistant voices of marginalized communities. This suppression can have serious implications, not only for the visibility of Black Twitter, but for its role as a space for cultural production and resistance.

Black Twitter Black Twitter has become a critical space for cultural production, activism, and resistance. Scholars like Brock (2012) and Williams (2016) frame Black Twitter as a cultural hub where linguistic creativity and shared experiences converge. Black Twitter has served as a site of political resistance and collective action, with movements like #BlackLivesMatter using the platform to mobilize and organize. Harlow and Benbrook (2019) and Jones (2013) explore how Black Twitter acts as both a space for social commentary and a platform for social empowerment. Through the use of humor, wordplay, and shared cultural knowledge, Black Twitter users navigate racial and social dynamics in ways that mainstream platforms may overlook or suppress.

Black Twitter also plays a key role in counteracting the marginalization of Black voices. It is a space where users assert their identity, engage in cultural expression, and create solidarity around shared experiences. Through these activities, Black Twitter serves as both a virtual community and a site of resistance against the algorithms and systems that attempt to control or suppress these voices.

Ethical Considerations Researching communities like Black Twitter raises critical ethical questions, particularly when the researcher's role is that of both observer and participant. Haraway (1988) critiques the notion of objectivity, emphasizing that claims of neutrality often serve to dismiss the voices of marginalized groups. She argues that instead of striving for an unattainable objectivity, researchers should embrace positionality, acknowledging their own biases and the ways in which their research practices can influence the communities they study. This aligns with Harrison (2016) argument that the perspectives of marginalized groups must be central to research and theory-building.

Ethical concerns in studying Black Twitter are compounded by the dynamics of digital research. Constable (2012) discusses the challenges of building trust in online spaces, particularly when the researcher's presence can disrupt the natural flow of community dynamics. Trust and authenticity are crucial when engaging with marginalized communities, and researchers must be mindful of their role in preserving the integrity of the communities they study.

Methodology for Ethical Checklist

In this section we discuss the various axes of ethical concerns and considerations that researchers should take when studying marginalized digital communities including positionality, data sharing, anonymity, and aggregation, the inclusion of community in the research pipeline, and observation of community discourse and practices. We will then contextualize each axes in the context of researching Black Twitter. Within this section we attempt to answer the question: What ethical challenges arise when studying marginalized online communities? At the end of the section, we pose a ethics checklist of actionable items that we encourage researchers to take in order to better protect studied communities.

Positionality of Researcher Fuchs (2018) highlights how data ownership becomes complex when individuals share content in public forums. Black Twitter, despite being hosted on a public platform, functions as a relatively enclosed community defined by shared cultural practices, followed accounts, and specific content. Without dedicated hashtags or explicit demarcated spaces, Black Twitter's boundaries are subtler, making it difficult for outsiders to identify its members and content accurately.

As a researcher and participant within this virtual space of Black Twitter, I [1] recognize the exclusivity and privacy that characterizes the community. This raises a critical question: Do I, or more generally Black researchers, have the right to research it? The answer is not straightforward. While Black researchers like myself seek to highlight the importance, influence, and power of Black Twitter, we must also acknowledge the potential costs. Many works infer that marginalized communities are more comfortable working with in-community members due to the distrust and extractive nature of out-group members as well as an overall distrust of research (Lee and Rich 2021; George, Duran, and Norris 2014). Dr. Cindy Peltier, a associate professor specializing in Indigenous and community-based research methods, shared in a 2021 Wired article discussing Indigenous data sovereignty and the study of Indigenous communities that "folks would come in and take information and then publish whatever they wanted without ever consulting the community" (Huckins 2021). Klassen and Fiesler (2022) explains that, "[s]everal participants expressed that White people doing research on Black Twitter ... seemed colonizing" and one of their participants noted that "[they] don't mind Black people in the academic world sort of reflecting and research-

[1] "I" is in reference to the first author.

ing these experiences because [they] think, on some levels, it's a shared experience. And [they] think those voices need to be heard in those areas" (pg 7). These perspectives underscore that positionality matters. Researchers within the community are better positioned to engage ethically whilst still approaching the work with humility, sensitivity, and awareness of impact.

Positionality refers to the idea that our situated knowledge, shaped by our identities, lived experiences, and sociocultural backgrounds, influences our perspectives, assumptions, and decision-making processes (Rose 1997; Secules et al. 2021; Yip 2024). For both in-group and out-group community members, researchers must consistently assess how their positionality and identity, in the context of the work being conducted, shape the decisions they make throughout the research process.

As researchers, we must ask ourselves: Why do we care about the work we do? What motivates us to pursue it? And how does our identity influence the lens through which we interpret this work?

This self-reflection is especially important for research that directly engages with or impacts communities, including work on fairness, ethics, and the deployment of machine learning systems. Yet even seemingly distant areas like information extraction or code generation must account for how data diversity, language accessibility, and inclusive design are shaped by the researcher's own positionality.

Regardless of community membership or the type of research you conduct, it is essential to examine our own biases and beliefs and how they shape the decisions we make and the research we pursue (Singh et al. 2025). Acknowledging one's positionality in relation to their work is a foundational step in responsible research.

Anonymity, Data Sharing, and Aggregation Anonymization of identifiable markers - such as handles, replies, and location - is a critical ethical step prior to analyzing or sharing data from marginalized communities. This is particularly important because Black Twitter often serves as a site for discourse and organizing, making vocal users vulnerable to suppression or targeting. Although users consented to share their content publicly under Twitter's guidelines, they did not consent to have their sometimes outlandish tweets be plastered on a research poster for out-group members to engage with out of context or without social understanding. The ability to trace a tweet or social media posting back to real identity of the author poses various threats to the separation of online and in-person existence, and for many individuals is a violation of their privacy. Beyond anonymization, ethical analysis of social media data also requires careful attention to how data is aggregated, as the ways we group and categorize users can significantly shape both the findings and their potential impact.

Aggregation of data must be approached with care. Researchers should avoid overgeneralization or drawing overly broad claims based on a small subset of data. While data can reveal trends, it is crucial to acknowledge this limitation and that the studied community is not homogeneous in thought and belief. Aggregation should be informed by existing social science literature to avoid harmful oversimplifications. For instance, grouping individuals from diverse Asian backgrounds under a single label like "Asian" ignores varying lived experiences and can lead to inaccurate conclusions. In many communities, there has already been a push for disaggregation of demographic groups. Studies have found that disaggregated methods have had explicit benefit to the individual communities and improved data quality and analysis when communities were assessed separately (Nguyen, Nguyen, and Nguyen 2014).

Parallel to this, within computer science, an increasing number of studies highlight the importance of preserving disagreement in subjective tasks such as hate speech detection or emotion classification (Sandri et al. 2023; Larimore et al. 2021). These works often note that even aggregating annotations across in-group and out-group members can obscure meaningful differences, as individuals within the same community may still express diverse and conflicting perspectives. By being more intentional in our aggregation practices and being more transparent about our methods, researchers can minimize the harm inflicted on the studied communities. In addition to careful aggregation, ethical research also demands thoughtful consideration of how data is shared, as the redistribution of content from marginalized communities can amplify harm if not handled with care.

Inclusion of the Community and Data Sovereignty We emphasize the need for participatory research. The inclusion of community members at every stage of the research pipeline is crucial. This extends beyond relying on researchers who identify as part of the community. Individual researchers bring their own biases and cannot fully represent the diversity of the community - they are not a representation of their community. Failing to involve the broader community risks undermining both the quality of the research and its relevance to those it seeks to represent.

Community involvement can take many forms. For example, researchers could partner with nonprofit organizations embedded in the community, involve representatives in research meetings, and incorporate their feedback at every step. Sharing major milestones with community members outside of academia and implementing their suggestions can further enhance the research. Employing in-group annotators to validate results ensures that the conclusions align with the community's perspectives. Using mixed methods, researchers can incorporate qualitative insights to complement quantitative findings, producing research that the community values and supports.

Furthermore, the inclusion of Indigenous data sovereignty principles and data practices should be utilized to ensure self-determination of studied marginalized virtual community. Walter et al. (2021) states that Indigenous data sovereignty "affirms the rights of Indigenous Peoples to determine the means of collection, access, analysis, interpretation, management, dissemination and re-use of data pertaining to the Indigenous Peoples from whom it has been derived, or to whom it relates". It is the idea that data extracted from these indigenous communities are cultural arti-

facts. We can apply similar principles for marginalized virtual communities in which the data and content they craft and create are culturally specific relics that embody social, political, and community norms. These relics thus need to be protected and managed by the communities they are extracted from.

Platform Capitalism Platform capitalism refers to the economic beneficiary structure that underlies many digital platforms, that is the commodification of user behavior and generated data for profit (Pasquale 2016; Srnicek 2017). Rather than finding value in goods and services, platforms like Twitter extract value from social interactions on their platforms. Under this model, platforms do not merely host communities, they exploit them. Spaces like Black Twitter are harvested by researchers, marketers, advertisers, and the platforms themselves for the rich cultural linguistic data that is labored by the users without consent. This creates a dynamic of data extraction, where Black cultural production fuels platform engagement and profit, but the communities that generate it are rarely acknowledged or materially supported. Black users have consistently been overrepresented on Twitter making up about 25% of U.S. users by 2010 despite only being about 13% of the general population (Heussner 2010). Twitter has generated significant revenue from this user base, capitalizing on trends and cultural conversations without investing in or protecting the community itself.

This disproportionately impacts marginalized digital communities, extracting and commodifying their lived experiences and thoughts, while over-moderating users when it seems inconvenient for the platform and their needs while allowing for racism and misogynoir to infiltrate the same safe spaces that these platforms claim to care and profit from (Noble 2018). In recent years, platforms like Twitter have failed to moderate the rise of anti-Black, Nazi, and racist content, contributing to a mass disengagement and exodus from Black Twitter (Dwoskin 2023; Siddiqui and Merrill 2023; Faverio 2023). This reflects a broader pattern: platforms capitalize on communities they did not build or sustain, and whose survival is often jeopardized by the very policies that make the platforms profitable.

Researchers, too, can inadvertently reproduce this extractive logic by relying on data produced by marginalized communities without critically examining the power structures that make such data accessible. Ethical research must move beyond checklists and surface-level safeguards to engage with the deeper question of how to disincentivize extraction, both in platform design and in research practice.

The Checklist

By adopting these considerations, researchers can reduce potential harms associated with studying marginalized and oppressed virtual communities. This ethical checklist prioritizes privacy, transparency, accountability, and meaningful community participation, ensuring that the research amplifies the voices of the studied community without exposing them to risk. In doing so, researchers can conduct research that not only meets academic standards but also earns the trust and respect of the communities studied. [2]

1. Aggregation should be informed by established social science work, especially around features such as race and ethnicity (Schwabish and Feng 2022) and should push against standard approaches such as majority-vote by considering the context and reasoning behind the aggregation step (Fleisig et al. 2024).
2. Do not assume demographic features of a user based on emoji use, geo-location information, or use of dialectal language unless explicitly specified by user; use continuous metric analysis such as automatic dialect density (Washington et al. 2018; Johnson et al. 2022) to better understand population features without making explicit ties to a group (i.e. even if someone uses African-American Vernacular English online, does not mean they are African-American) (Wong 2019; Henning 2025).
3. Always annonymize data (e.g. handles, geo-location, published date and time) through approaches like masking (i.e. @this_is_my_handle$\rightarrow$ @username) or deletion of certain features.
4. Regulate distribution of data; put data behind access walls through request forms and justification for potential use of data to keep bad actors (and even well-intentioned actors with poor discernment of use case) from misusing data.
5. When studying closed community groups (e.g. private Facebook groups or private community threads on Twitter), shift ownership and control of the data to the leads of these communities empowering them to decide the dissemination and storage of data; draw strategies from existing Indigenous data sovereignty researchers work (Rainie et al. 2019; Walter et al. 2021).
6. Leverage both quantitative and qualitative approaches as online and offline communities influence one another as we gain nuance and experiential understanding through qualitative analysis; conduct interviews and host surveys of the studied population to include community voice.
7. Assess your positionality throughout the research process and document reasoning of decisions made and how your understanding influences these decisions; include positionality statement; answer the questions: Why do we care about the work we do? What motivates us to pursue it? And how does our identity influence the lens through which we interpret this work?

While this checklist is not an exhaustive guide to resisting extractive research practices, it offers an important starting point. That said, many of the steps outlined may be difficult to implement without institutional support or access to adequate resources. Institutional buy-in can be difficult to secure, but top-down strategies can help. These often involve framing methods to align with institutional values, emphasizing the utility of community-grounded approaches for developing more effective and applicable inter-

[2]**Note**: Throughout the case studies and discussion, we note which item on the checklist is being used by referencing it in the format (**Item 1**).

ventions, or appealing to the tried and true argument that inclusive socio-technical systems are ultimately more robust. But these strategies come with trade-offs. To gain traction, they often require flattening the social complexity at the core of this work, translating deeply contextual and relational research into simplified narratives that institutions find legible (Ovalle et al. 2023). This oversimplification is not just a strategic concession; it can be a harm. It reduces communities to tools for optimization, positioning inclusion as a means to improve model accuracy rather than as a response to structural injustice. In doing so, it risks erasing the ongoing harms these communities face and creates space for institutions to declare "success" once certain performance metrics are met, regardless of whether meaningful change has occurred. Nonetheless, even with these approaches, institutional norms and broader community standards often carry more weight in decision-making than the merits of these methods alone, particularly in the eyes of funding institutions. Therefore to address this, we highlight several items on the checklist that are both achievable and should be considered a bare minimum for all researchers. Items 1, 2, and 3, focused on data processing and analysis, are designed to be accessible regardless of a researcher's institutional backing. Item 7, which calls for proactive, introspective engagement through a positionality statement, similarly requires minimal resources. While positionality statements can raise concerns around page limits, we recommend including them within the Ethical Considerations section of papers, which should ideally be exempt from page count restrictions. Incorporating these four items, though modest in scope, lays the groundwork for more ethical engagement and helps foster a broader culture of accountability and resistance to extractive research practices.

Case Studies

We perform two case studies leveraging our proposed checklist to study Black Twitter through the lens of the n-word, a reclaimed slur within the Black community.

The n-word, a historic slur used against Black Americans as early as the 1600, has roots in slavery and the Jim Crow era. It has been used to diminish the character and existence of Black Americans within the United States for centuries. However, the word, re-imagined, was introduced into mainstream media via Hip Hop in the late 1980s (Asim 2007). Since then, the word has taken on two distinct forms, the first being the original derogatory use, and the other a term of comradery and membership mainly used among Black people in the United States. However due to the continual contentious use of the word, work studying the actual reclaimed-ness of the word differs between linguists, social scientists, and non-academics (Smith 2019; Rahman 2012). This contention allows for the study of usage and learned norms around the word to be fruitful and rich in discovery around the impact of environment on acceptability of cultural features.[3]

[3]While this study focuses on the use of the n-word in the context of Black Twitter and its users, it does not seek to essentialize the term as the defining feature of Black culture, particularly not Black culture as it manifests within Black Twitter. Black Twitter represents a dynamic and multifaceted digital community with a long history of political activism, in-community accountability, cultural commentary, and community building. From driving national conversations to shaping pop culture, its influence extends far beyond any single linguistic practice. This work aims to engage with one aspect of Black discourse without reducing the richness and complexity of Black Twitter to the usage of the n-word alone.

Demographic Feature	N	%
Ethnicity		
Caribbean-American, Jamaican, Black	2	16.67
Black Southeast-African	1	8.33
African-American, Black	1	8.33
Black, Ugandan/American	1	8.33
Black	3	25.00
African-American	1	8.33
African-American, Black	2	16.67
African-American, Black, Caribbean-American, Jamaican, Puerto Rican	1	8.33
Age		
24	2	16.67
22	2	16.67
26	1	8.33
25	2	16.67
23	3	25.00
27	2	16.67
Are you Black?		
Yes	12	100.00
No	0	0.00
Are you a woman?		
Yes	6	50.00
No	6	50.00

Table 1: Demographic features of the survey respondents.

We use these case studies to 1) qualitatively investigate and learn how reclaimed word usage varies within the Black community and discuss the bounds of usage of the word and 2) quantitatively assess how the usage of this words and different perturbations of this word online signal in-group vs out-group usage and how this community finds ways to continue usage despite biased automatic toxicity and hate speech classification moderation systems online. Through these two case studies, we gain insight into word use and community opinions while respecting the communities privacy as well as aligning with ethical concerns. Additionally, we hope to answer the question: How do algorithms affect the visibility and interactions within Black Twitter?

Qualitative Case Study on Observing Discourse, Language, and Practices

Researchers use virtual ethnography to observe the cultural practices, language use, and communal norms within Black Twitter. By focusing on the digital communication strategies employed by Black Twitter users, it highlights how race, identity, and culture are performed in a virtual space. Florini (2014) examines how the tradition of "signifyin'", a core aspect of Black American cultural expression, is adapted and enacted on Twitter. This form of cultural performance, along

with the use of reclaimed language, marking, and loud-talking, and other communal linguistic features, allows users to navigate and assert their identity in a space that might otherwise obscure their racial and cultural background (Rickford 2015). Building on prior work that has employed this approach to study online communities, this work adapts several of these informal methods and principles.

With this context in mind, the following study examines how the lived experiences of social media users shape their use of the n-word as a reclaimed term and how that translates into online spaces. This study underscores the idea that no community is a monolithic entity. This work also investigate participants' expectations and perceptions regarding the word's use within in-group and out-group communities, as well as how in-group membership is defined for the n-word.

To conduct this study, a survey was hosted through Google Forms, gathering demographic information and asking participants a series of questions about their use of the word. This survey is part of a larger project exploring how members of Black Twitter perceive toxicity surrounding the n-word and how content moderation systems can better support marginalized communities. Twelve participants who identify as Black social media and Twitter users completed the form. The unique demographic features, with emails removed (**Item 3**), of our subpopulation are shown in Table 1. In the context of the larger Black population, a subsample of 12 participants is not representative, limiting our ability to generalize these findings. Therefore, we focus our analysis on insights drawn from this specific subset of surveyed individuals, rather than making broader claims about the wider community.

Our survey revealed several key insights. The first stage of analysis focused on identifying which community the reclaimed word is most closely associated with. Participants shared differing views on whether the n-word belongs exclusively to the Black African-American community or whether its usage extends to the broader African Diaspora, given its connotations of camaraderie. Some argued that, because the slur was historically weaponized against African Americans, its reclamation should remain specific to African Americans or second-generation Black individuals in the United States who have experienced oppression within that context. The question of ownership over the n-word, and by extension many elements of Black culture, has long been a contentious issue within the Black community. On Black Twitter, this debate frequently surfaces in discussions about appropriation and the erasure of Black African-American culture amid the celebration of broader Black culture. This dynamic reflects just one dimension of the larger conversations occurring within Black Twitter.

Another set of questions focused on personal and collective usage of the word. 41.7% of the respondents reported that they do not personally use the reclaimed word. However, 91.7% agreed that in-group individuals should use it if they choose to, and all participants acknowledged knowing in-group members who use the word. Additionally, 83.3% of participants believed that out-group individuals should never use the word, though 58.3% indicated they knew out-group individuals who have used it. These findings offer meaningful insights into the dynamics of reclaimed language in online communities and participants' comfort levels with engaging with content that features such language. Even if a content moderation system could accurately distinguish between in-group and out-group members, should it allow all in-group members to use and view content that includes their community's reclaimed language?

We also examined the word's use in reference to and in spaces with others. 70% of participants agreed that the word should not be used in reference to out-group individuals.[4] Participants were divided, however, on whether it should be used in the presence of out-group individuals. One participant noted in the additional context section that the n-word "should only be used by and in the company of Black folks." When applied to Black Twitter, this perspective raises questions about who engages with and participates in content within Black Twitter. While there was no majority consensus among participants, the suggestion of limiting its use to Black Twitter members highlights broader questions about how Black Twitter can remain predominantly or exclusively Black on a platform that does not support private groups or spaces.

Through this small subset of participants, the diversity of opinions on what might initially appear to be a straightforward and widely agreed-upon topic, the use of the n-word, was explored. Studies like this, combining small sample sizes with mixed methods, offer valuable insights into a community's dynamics without requiring full access to its entire population, with the understanding that such findings do not represent the community as a whole.

Quantitative Case Studies on Algorithmic Impact

Case studies are a common method for assessing specific features of marginalized online communities. Studies such as Zhao et al. (2017) on AI biases and Dorn et al. (2024) on harmful speech detection highlight how algorithmic biases disproportionately affect marginalized users. Algorithmic case studies, such as audits, that assess the interplay between algorithmic soico-technical systems and the social contexts they are employed in allow us to study the impact these systems have (Christin 2020). These studies reveal the limitations of current algorithmic and content moderation systems and underscore the need for more equitable content moderation practices.

In this case study, we explore a workaround for term biases in existing content moderation systems, focusing on reclaimed language through what Calhoun and Fawcett (2023) describes as linguistic self-censorship. Originally defined in the context of TikTok and other video-based platforms, linguistic self-censorship refers to

> any instance of a social media creator intentionally changing their linguistic practices to avoid using a specific word or phrase, either because of potential risk for actual censorship through algorithmic enforcement of community guidelines or as a form of mimetic language play (Calhoun and Fawcett 2023, p. 2).

[4]Note that only 10 respondents answered this question.

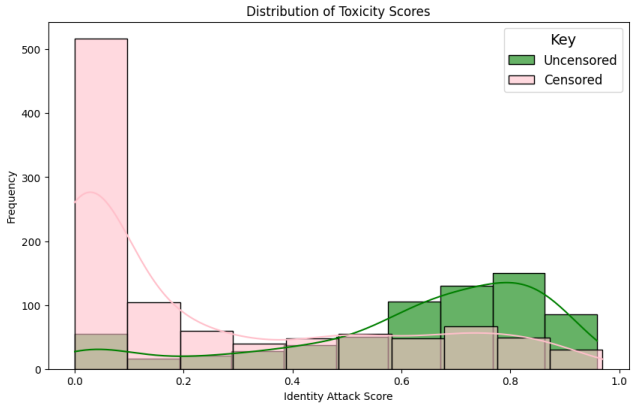

Figure 1: Distribution of identity attack scores from the Detoxify model comparing tweets using censored vs uncensored forms of the n-word.

While this concept was initially applied to video platforms, it can be generalized to other forms of social media, where users adjust their language to evade suppression by content moderation algorithms. This study examines linguistic self-censorship of the n-word, specifically intentional misspellings or character substitutions, and its effects on toxicity classification models.

Term Bias Previous research has demonstrated that content moderation systems exhibit term bias. This refers to the disproportionate weight given to certain identity markers (e.g. 'Black', 'Jewish'), slurs (e.g. n-word, f-word, k-word), and expletives (e.g. 'fuck', 'shit', 'ass') in text classification, regardless of the content or intent of their usage (ElSherief et al. 2021; van Aken et al. 2018; Vidgen et al. 2019).

To counter this, many users employ character perturbations, such as replacing vowels with symbols (e.g. asterisks or ampersands) or substituting letters with similar numbers (e.g. f*ck, sh*t, b1tch, wh0re). These strategies make it more difficult for content moderation systems to flag and remove posts. This approach is especially prevalent in marginalized online communities, where reclaimed language is a significant part of online vernacular.

Methodology This analysis focused on tweets containing both the linguistically self censored and uncensored version of the n-word (including its full and shortened form which are often used as reclaimed language). Tweets were collected using Twitter's Pro API on June 17th, August 2nd, September 16th, and September 19th, 2024. The queries targeted both singular and plural forms of the n-word with -er and -a endings, as well as a perturbed variant in which the letter "i" was replaced with an asterisk (*). From these queries, 2,780 tweets were obtained. In this study, we adhered to this social norm and only used the conventionally accepted censorship of the n-word.

After text processing, including removing URLs, masking Twitter handles, converting emojis to textual descriptions, and removing duplicate tweets (**Item 3**), the dataset was reduced to 1,695 tweets. Each tweet was then categorized based on whether it contained a censored or uncensored version of the n-word. Of these, 1,018 tweets used a censored version, while 677 used an uncensored version. To evaluate toxicity, we used the Detoxify model, a toxic comment classifier developed by Unitary(Hanu and Unitary team 2020). Detoxify was trained to predict toxic comments using data from three Jigsaw challenges: Toxic Comment Classification Challenge, Unintended Bias in Toxicity Classification, and Multilingual Toxic Comment Classification. It produces scores for several categories, including toxic, severe toxic, obscene, threat, insult, and identity threat. The tweets were processed through the Detoxify model, focusing on the identity attacks toxicity score, as the use of slurs is central to this analysis. As seen in Figure 1, we visualized the results using a histogram with the following parameters: bins = 10, kde=True, alpha =0.6.

Findings Our analysis and results in Figure 1 reveal a clear trend: tweets using self-censorship generally have lower toxicity scores for identity attacks compared to tweets using the uncensored version. The findings suggest two things. First, it suggests a distinction between in-group and out-group usage. Users who use the uncensored version of the n-word may belong to out-groups and often employ the term in derogatory or hateful contexts. In contrast, those using the censored version, following the common Black Twitter norm of censoring vowels rather than consonants, are more likely to be in-group members using the term in a positive, reclaimed way, connoting camaraderie and shared identity.

Additionally, we found that tweets using the uncensored n-word received higher identity attack scores from Detoxify compared to those using the censored version. Several confounding factors could influence this result, such as whether the tweet's context was hateful or if it was mainly influenced by term bias. To address these confounding factors, we performed a perturbation where we swapped the censored and uncensored versions of the term. This allowed us to control for the tweet's context and specifically examine term bias. Figure 2 presents two violin plots, with each plot corresponding to the form of the n-word used in the original tweet and illustrating the difference between the uncensored and censored identity attack scores. This figure indicates that self-censorship generally reduces the attack score significantly, with an average score difference of 0.307 between the original censored and uncensored versions (0.204 for censored and 0.462 for uncensored). This suggests that Detoxify relies heavily on the presence of the uncensored term itself, rather than on the surrounding context, when making toxicity classifications. In contrast, tweets that use censored versions of the n-word are often rated as less toxic, even when the underlying sentiment remains unchanged. This highlights a critical limitation of current toxicity detection models: they prioritize surface-level term matching over contextual understanding. As a result, linguistic self-censorship becomes a necessary strategy for marginalized users not only to evade unjust suppression but also to preserve the in-group usage of reclaimed language that fosters identity, solidarity, and cultural expression within Black

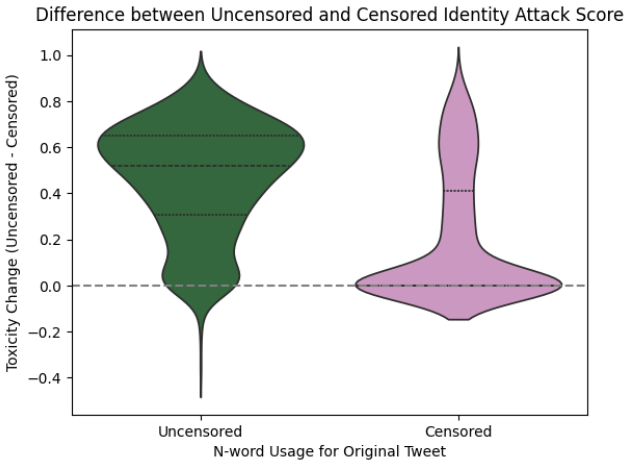

Figure 2: Distribution of the difference of Uncensored - Censored identity attack scores from the Detoxify model comparing tweets using censored vs uncensored forms of the n-word from the original text.

Twitter.

This small case study demonstrates how marginalized online communities navigate and resist the existing limitations of content moderation systems. By leveraging these strategies like linguistic self-censorship, these communities can effectively push back against algorithmic suppression to ensure their content remains visible and more generally ensure an existence of their virtual community. Additionally, we were able to conduct this analysis without exposing the community or specific participants to harm and without engaging in uninformed aggregation.

Discussion

Case Study Results This study addresses these central questions posed in the introduction by proposing an ethical checklist for researching marginalized virtual communities and applying it to two case studies informed by the checklist. First we ask what ethical challenges arise when studying marginalized virtual communities? To answer this, an ethical framework for conducting ethnography in virtual spaces and communities is presented, emphasizing privacy and transparency, proposing harm-reduction approaches for data sharing, and outlining meaningful ways to include the studied community throughout the research process. These methods are then applied to subsequent studies, ensuring ethical considerations remained central to the research.

Second we ask, how can we conceptualize Black Twitter as a virtual community? Using mixed-method approaches, we probed the virtual presence of marginalized communities, focusing on small yet significant features such as the comfort and acceptability of reclaimed language (i.e. the n-word). This analysis reveals the nuanced ways in which language fosters identity and belonging, providing insight into Black social media users and a culturally rich discussion within the Black Diaspora. Specifically, the discussion of the ownership of the n-word is such a complex one, shown by the drastic variance in responses. While all participants agreed that the term belongs to — and should remain within — the Black-identifying community, the more compelling question lies in how "Black" is defined. The word carries immovable ties to slavery and a deep-rooted history in the United States. However, with the increasing number of Black immigrants entering the country, the boundaries of Black identity within the U.S. are shifting. This shift is reflected in some of the more specific responses, such as those that reference the number of generations someone's family has held American citizenship. These responses offer a window into how deeply layered and interconnected questions of identity can be. More importantly, they prompt us to consider how these complexities are translated into digital spaces—and how the idea of a global online community further influences the already nuanced and sensitive bonds of Black identity.

Finally we ask, how do algorithms influence and affect visibility and interactions within Black Twitter? Through an algorithmic case study, we observe online trends in linguistic adaptations employed by marginalized communities to resist algorithmic suppression and ensure their survival online. The case study on linguistic self-censorship demonstrates one method of navigating the limitations of content moderation systems, offering a small yet meaningful glimpse into how algorithms shape digital interactions and visibility. Our results highlight the effectiveness of linguistic self-censorship, demonstrated by the substantial decrease in identity attack scores for tweets that originally contained the uncensored n-word once they are censored. This suggests that automated systems place significant and potentially biased weight on the presence of the n-word itself. In contrast, tweets that were already censored show a much smaller change in score, indicating that their usage is less likely to be interpreted as hateful. While we are unable to analyze these findings by user demographics, the patterns suggest that many of the originally censored tweets are likely examples of reclamatory use rather than hate speech (**Item 6**).

Applying the Checklist Throughout both case studies, we applied several of the recommended practices from the proposed ethical checklist. I explicitly reflected on my positionality as both a member of the Black community and an active participant in Black Twitter, acknowledging how this shaped my assumptions and interpretations (**Item 7**). For example, I initially assumed that tweets containing the uncensored n-word, excluding explicitly derogatory uses ending in "-er", were written by individuals outside the Black community. This assumption was based on a biased belief that Black Twitter users would not use the term in contexts aligned with hate speech (**Item 2**). This assumption was flawed not only because the racial identity of tweet authors was unknown, but also because it falsely implied moral innocence on the part of in-group users. We also prioritized anonymization: before analysis, we removed any potential PII such as email addresses and user handles to protect individual identities. Finally, due to the sensitive nature of the dataset, we do not plan to release it publicly (**Item 4**). Keeping the data internal helps prevent misuse and preserves its intended pur-

pose which was for evaluating how moderation systems handle reclaimed slurs rather than enabling unintended applications. Using this checklist compelled me to more deeply consider the broader ethical context of the research.

Conclusion

This study introduces an ethical, community-centered checklist for researching marginalized online communities, drawing from virtual ethnographic methods and Indigenous data practices. Unlike broader ethical frameworks, this checklist offers more actionable and context-specific guidance tailored to virtual communities. We apply the checklist in a mixed-methods study focused on Black Twitter and the use of the n-word, examining the diverse and often conflicting perspectives within the community. Our findings highlight not only the lack of consensus around the word's usage but also the creative strategies users employ, such as linguistic self-censorship, to navigate and resist biased toxicity classification systems.

This work takes a meaningful step toward more accountable and community-informed research practices, offering a broadly applicable and adaptable checklist that can guide ethical engagement with marginalized virtual communities. While rooted in a specific platform and linguistic phenomenon, the checklist we present is intentionally designed to be extensible and relevant across diverse socio-technical environments. We see this not as a constraint, but as a base for future scholarship.

To that point, we invite future work to explore how this checklist functions across a broader range of platforms with differing architectures, such as those with varying moderation policies, levels of anonymity, and forms of algorithmic amplification. Our approach also opens the door for participatory design methodologies, where community members take an active role in shaping and refining ethical guidelines. Moreover, this work lays a foundation for re-imagining platform governance itself, encouraging more collaborative, user-facing systems that move beyond one-size-fits-all solutions and better support the autonomy and cultural nuance of marginalized communities. We hope this work inspires further community-centered research and design that pushes the boundaries of how we ethically and effectively engage with digital communities.

Limitations

One of the major limitations of the study is the subpopulation size for the qualitative case study. A subpopulation of only 12 individuals is not representative of an entire population, which was acknowledged within our work. Additionally we are not able to draw any conclusions about the beliefs of the population as one of the goals of the case study was the highlight the variance in opinion. With a greater sample size, we would be able to gather more information on varying opinions within the community. Another major limitation is the bounds in which we discuss Black Twitter. As Black Twitter is not always ascribed to one hashtag such as #BlackTwitter, unless linked to a specific movement or topic such as #BlackLivesMatter or #ThanksgivingClapBack, we are unable to specifically gather tweets and post from directly within Black Twitter. This work positions the conversation of Black Twitter members' opinions of the n-word usage and the actual usage of the n-word online to try to highlight approaches known and leveraged within the Black Twitter community.

Ethical Consideration

Positionality Statement. The first author of this study is a Black woman who regularly participates in Black Twitter and other Black virtual communities across social media platforms. This work is informed by her engagement with scholarship and content that often overlooks the safety and privacy of Black Twitter in pursuit of publicizable or attention-grabbing research. She is additionally a daughter of Caribbean-immigrants, and the first American born individual in her immediate family. Her cultural context and community does influence her understanding of reclaimed language more generally and especially the n-word.

Dataset Distribution. This study used a subset of roughly 2.7 thousand tweets containing the various forms of the n-word from another project that collected tweets for a variety of reclaimed languages. Due to potential misuse, we will not be sharing the dataset and will keep it closed to the current research team that helped curate dataset as to reduce harms.

Participant Survey. This study is covered under the IRB approval of the project used to curate the dataset. That project went through IRB approval and received IRB exemption. Any Personally Identifiable Information from the survey collection process has been de-identified to protect the privacy of the respondents.

Acknowledgments

The authors would like to thank Arjun Subramonian, Rebecca Pattichis, and Ashima Suvarna for their constant support and feedback on this paper and case studies. This work was funded in part by the Eugene V. Cota-Robles Fellowship to Christina Chance and was done in part while Kai-Wei Chang was visiting the Simons Institute for the Theory of Computing.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *ProPublica*.

Arora, A.; Barrett, M.; Lee, E.; Oborn, E.; and Prince, K. 2023. Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization.

Asim, J. 2007. *Nigger vs. Nigga*, 212–257. Houghton Mifflin Harcourt.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Brock, A. 2012. From the Blackhand Side: Twitter as a Cultural Conversation. *Journal of Broadcasting & Electronic Media*, 56(4): 529–549.

Calhoun, K.; and Fawcett, A. 2023. "They Edited Out her Nip Nops": Linguistic Innovation as Textual Censorship Avoidance on TikTok. *Language@Internet*, 21: 1–30.

Christin, A. 2020. The ethnographer and the algorithm: beyond the black box. *Theory and Society*, 49(5): 897–918.

Colón Vargas, N. 2024. Exploiting the margin: How capitalism fuels AI at the expense of minoritized groups. *AI and Ethics*, 1–6.

Constable, N. 2012. Correspondence Marriages, Imagined Virtual Communities, and Countererotics on the Internet. *Media, erotics, and transnational Asia*, 111–138.

Dev, S.; Monajatipoor, M.; Ovalle, A.; Subramonian, A.; Phillips, J.; and Chang, K.-W. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1968–1994. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Dorn, R.; Kezar, L.; Morstatter, F.; and Lerman, K. 2024. Harmful Speech Detection by Language Models Exhibits Gender-Queer Dialect Bias. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712227.

Dwoskin, E. 2023. Fleeing Elon Musk's X, the quest to recreate 'black twitter'. *The Washington Post*, 6.

Ellis, D.; Oldridge, R.; and Vasconcelos, A. 2004. Community and virtual community. *Annual Review of Information Science and Technology*, 38(1): 145–186.

ElSherief, M.; Ziems, C.; Muchlinski, D.; Anupindi, V.; Seybolt, J.; De Choudhury, M.; and Yang, D. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 345–363. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Elwood, S.; and and, A. L. 2018. Feminist digital geographies. *Gender, Place & Culture*, 25(5): 629–644.

Farahani, M.; and Ghasemi, G. 2024. Artificial intelligence and inequality: Challenges and opportunities. *Int. J. Innov. Educ*, 9: 78–99.

Farayola, M. M.; Tal, I.; Malika, B.; Saber, T.; and Connolly, R. 2023. Fairness of AI in Predicting the Risk of Recidivism: Review and Phase Mapping of AI Fairness Techniques. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, ARES '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400707728.

Faverio, M. 2023. Majority of US Twitter users say they've taken a break from the platform in the past year. *Pew Research Center*, 17.

Fleisig, E.; Blodgett, S. L.; Klein, D.; and Talat, Z. 2024. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2279–2292. Mexico City, Mexico: Association for Computational Linguistics.

Florini, S. 2014. Tweets, Tweeps, and Signifyin': Communication and Cultural Performance on "Black Twitter". *Television & New Media*, 15(3): 223–237.

Fuchs, C. 2018. 'Dear Mr. Neo-Nazi, can you please give me your informed consent so that I can quote your fascist tweet?': Questions of social media research ethics in online ideology critique. In *The Routledge companion to media and activism*, 385–394. Routledge.

George, S.; Duran, N.; and Norris, K. 2014. A Systematic Review of Barriers and Facilitators to Minority Research Participation Among African Americans, Latinos, Asian Americans, and Pacific Islanders. *American Journal of Public Health*, 104(2): e16–e31. PMID: 24328648.

Hair, N.; and Clark, M. 2007. The ethical dilemmas and challenges of ethnographic research in electronic communities. *International Journal of Market Research*, 49(6): 1–13.

Haj Ahmad, N.; Stigholt, L.; Penzenstadler, B.; and Duboc, L. 2025. Ai Systems' Negative Social Impact and Factors. *Linnea and Penzenstadler, Birgit and Duboc, Leticia, Ai Systems' Negative Social Impact and Factors*.

Hanu, L.; and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Haraway, D. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3): 575–599.

Harlow, S.; and Benbrook, A. 2019. How #Blacklivesmatter: exploring the role of hip-hop celebrities in constructing racial identity on Black Twitter. *Information, Communication & Society*, 22(3): 352–368.

Harrison, F. V. 2016. Theorizing in ex-centric sites. *Anthropological Theory*, 16(2-3): 160–176.

Henning, T. M. 2025. Digital Blackface and Its Argumentative Implications. *Ethical Theory and Moral Practice*, 1–20.

Heussner, K. 2010. Is Twitter disproportionately popular among Black users.

Huckins, G. 2021. For Marginalized Groups, Being Studied Can Be a Burden. *Wired*.

Johnson, A.; Everson, K.; Ravi, V.; Gladney, A.; Ostendorf, M.; and Alwan, A. 2022. Automatic dialect density estimation for african american english. *arXiv preprint arXiv:2204.00967*.

Jones, F. 2013. Is Twitter the underground railroad of activism? *Salon*.

Katyal, S. K. 2022. Democracy & Distrust in an Era of Artificial Intelligence. *Daedalus*, 151(2): 322–334.

Klassen, S.; and Fiesler, C. 2022. "This Isn't Your Data, Friend": Black Twitter as a Case Study on Research

Ethics for Public Data. *Social Media + Society*, 8(4): 20563051221144317.

Knowles, B.; Fledderjohann, J.; Richards, J. T.; and Varshney, K. R. 2023. Trustworthy AI and the Logics of Intersectional Resistance. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 172–182. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J. R.; Jurafsky, D.; and Goel, S. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689.

Kuntsman, A. 1970. Cyberethnography as home-work. *Anthropology Matters*, 6.

Larimore, S.; Kennedy, I.; Haskett, B.; and Arseniev-Koehler, A. 2021. Reconsidering Annotator Disagreement about Racist Language: Noise or Signal? In Ku, L.-W.; and Li, C.-T., eds., *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 81–90. Online: Association for Computational Linguistics.

Latham, A.; and Crockett, K. 2024. Towards Trustworthy AI: Raising awareness in marginalized communities. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

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*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Lee, N.; Jung, C.; and Oh, A. 2023. Hate Speech Classifiers are Culturally Insensitive. In Dev, S.; Prabhakaran, V.; Adelani, D. I.; Hovy, D.; and Benotti, L., eds., *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 35–46. Dubrovnik, Croatia: Association for Computational Linguistics.

McElroy, E.; and Vergerio, M. 2022. Automating gentrification: Landlord technologies and housing justice organizing in New York City homes. *Environment and Planning D: Society and Space*, 40(4): 607–626.

McInroy, L. B. 2016. Pitfalls, Potentials, and Ethics of Online Survey Research: LGBTQ and Other Marginalized and Hard-to-Access Youths. *Social Work Research*, 40(2): 83–94.

Mhlambi, S.; and Tiribelli, S. 2023. Decolonizing AI ethics: Relational autonomy as a means to counter AI harms. *Topoi*, 42(3): 867–880.

Mott, C.; and and, D. C. 2021. Understanding how hatred persists: situating digital harassment in the long history of white supremacy. *Gender, Place & Culture*, 28(11): 1521–1540.

Nguyen, B.-M. D.; Nguyen, M. H.; and Nguyen, T.-L. K. 2014. Advancing the Asian American and Pacific Islander Data Quality Campaign: Data Disaggregation Practice and Policy. *Asian American Policy Review*, 24: 55–67. Name - New York University; Bureau of the Census; Department of Commerce; Coastline Community College; Copyright - Copyright Harvard Journal of African American Public Policy 2013/2014; Document feature - Graphs; ; Last updated - 2023-11-20; SubjectsTermNotLitGenreText - United States–US.

Noble, S. U. 2018. *Algorithms of oppression: How search engines reinforce racism*. New York university press.

Olson, L.; Guzmán, E.; and Kunneman, F. 2023. Along the Margins: Marginalized Communities' Ethical Concerns about Social Platforms. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, 71–82.

Ovalle, A.; Subramonian, A.; Gautam, V.; Gee, G.; and Chang, K.-W. 2023. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 496–511. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Parthasarathy, S.; and Katzman, J. 2024. Bringing communities in, achieving AI for all. *Issues in Science and Technology*, 10.

Pasquale, F. 2016. Two narratives of platform capitalism. *Yale L. & Pol'y Rev.*, 35: 309.

Rahman, J. 2012. The N Word: Its History and Use in the African American Community. *Journal of English Linguistics*, 40(2): 137–171.

Rainie, S. C.; Kukutai, T.; Walter, M.; Figueroa-Rodríguez, O. L.; Walker, J.; and Axelsson, P. 2019. *Indigenous data sovereignty*, chapter 21, 300–319. African Minds and the International Development Research Centre (IDRC).

Rheingold, H. 2000. *The virtual community: Homesteading on the electronic frontier*. The MIT Press. ISBN 9780262291415.

Rickford, J. 2015. African American Vernacular English in California: Over Four Decades of Vibrant Variationist Research. In Lanehart, S., ed., *The Oxford Handbook of African American Language*, chapter 15. Oxford: Oxford University Press.

Rogers, A.; Baldwin, T.; and Leins, K. 2021. 'Just What do You Think You're Doing, Dave?' A Checklist for Responsible Data Use in NLP. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4821–4833. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Rose, G. 1997. Situating knowledges: positionality, reflexivities and other tactics. *Progress in Human Geography*, 21(3): 305–320.

Sandri, M.; Leonardelli, E.; Tonelli, S.; and Jezek, E. 2023. Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2428–2441. Dubrovnik, Croatia: Association for Computational Linguistics.

Schwabish, J.; and Feng, A. 2022. Considerations for Ensuring Data Aggregation Is as Inclusive as Possible. Accessed: 2025-05-17.

Secules, S.; McCall, C.; Mejia, J. A.; Beebe, C.; Masters, A. S.; L. Sánchez-Peña, M.; and Svyantek, M. 2021. Positionality practices and dimensions of impact on equity research: A collaborative inquiry and call to the community. *Journal of Engineering Education*, 110(1): 19–43.

Siddiqui, F.; and Merrill, J. 2023. Elon Musk's Twitter pushes hate speech, extremist content into 'For You'pages. *Washington Post*.

Singh, A.; Dechant, M. J.; Patel, D.; Soubutts, E.; Barbareschi, G.; Ayobi, A.; and Newhouse, N. 2025. Exploring positionality in HCI: Perspectives, trends, and challenges. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Smith, H. L. 2019. Has nigga Been Reappropriated as a Term of Endearment?: A Qualitative and Quantitative Analysis. *American Speech*, 94(4): 420–477.

Srnicek, N. 2017. *Platform capitalism*. John Wiley & Sons.

Ungless, E. L.; Vitsakis, N.; Talat, Z.; Garforth, J.; Ross, B.; Onken, A.; Kasirzadeh, A.; and Birch, A. 2025. The Only Way is Ethics: A Guide to Ethical Research with Large Language Models. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 8992–9005. Abu Dhabi, UAE: Association for Computational Linguistics.

van Aken, B.; Risch, J.; Krestel, R.; and Löser, A. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In Fišer, D.; Huang, R.; Prabhakaran, V.; Voigt, R.; Waseem, Z.; and Wernimont, J., eds., *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 33–42. Brussels, Belgium: Association for Computational Linguistics.

Vidgen, B.; Harris, A.; Nguyen, D.; Tromble, R.; Hale, S.; and Margetts, H. 2019. Challenges and frontiers in abusive content detection. In Roberts, S. T.; Tetreault, J.; Prabhakaran, V.; and Waseem, Z., eds., *Proceedings of the Third Workshop on Abusive Language Online*, 80–93. Florence, Italy: Association for Computational Linguistics.

Walter, M.; Lovett, R.; Maher, B.; Williamson, B.; Prehn, J.; Bodkin-Andrews, G.; and Lee, V. 2021. Indigenous Data Sovereignty in the Era of Big Data and Open Data. *Australian Journal of Social Issues*, 56(2): 143–156.

Washington, J. A.; Branum-Martin, L.; Sun, C.; and Lee-James, R. 2018. The impact of dialect density on the growth of language and reading in African American children. *Language, speech, and hearing services in schools*, 49(2): 232–247.

White, G. 2019. What is Black Twitter and how is it changing the national conversation? Baylor Expert explains. *Media and Public Relations— Baylor University. Retrieved February*, 21: 2022.

Williams, A. 2016. On Thursdays we watch scandal: Communal viewing and Black Twitter. *Digital sociologies*, 273–293.

Wong, E. 2019. Digital blackface: How 21st century Internet language reinforces racism. *PhD Diss., UC Berkeley*.

Yip, S. Y. 2024. Positionality and reflexivity: negotiating insider-outsider positions within and across cultures. *International Journal of Research & Method in Education*, 47(3): 222–232.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. Copenhagen, Denmark: Association for Computational Linguistics.