INTERNET POLICY REVIEW
Journal on internet regulation

Volume 14 **Issue** 1

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Aspirational platform governance: How creators legitimise content moderation through accusations of bias

Blake Hallinan *Hebrew University of Jerusalem*
CJ Reynolds *Hebrew University of Jerusalem*
Yehonatan Kuperberg *Hebrew University of Jerusalem*
Omer Rothenstein *Hebrew University of Jerusalem*

DOI: https://doi.org/10.14763/2025.1.1829

Published: 31 March 2025
Received: 17 March 2024 **Accepted:** 23 October 2024

Funding: This research was funded by a grant from the Smart Institute at the Hebrew University of Jerusalem.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Hallinan, B., Reynolds, C., Kuperberg, Y., & Rothenstein, O. (2025). Aspirational platform governance: How creators legitimise content moderation through accusations of bias. *Internet Policy Review*, *14*(1). https://doi.org/10.14763/2025.1.1829

Keywords: Algorithmic bias, Content moderation, Platform governance, Creators, React videos

Abstract: While content moderation began as a solution to online abuse, it has increasingly been framed as a source of abuse by a diverse coalition of users, civil society organisations, and politicians concerned with platform bias. The resulting crisis of legitimacy has motivated interest in more participatory forms of governance, yet such approaches are difficult to scale on platforms that lack bounded communities and designated tools to support collective governance. Within this context, we use a high-profile debate surrounding bias and racism in content moderation on YouTube to investigate how creators engage in meta-moderation, the participatory evaluation of moderation decisions and policies. We conceptualise the conversation that plays out across a network of videos and comments as aspirational platform governance, or the desire to influence content moderation without established channels or guarantees of success. Through a content analysis of 115 videos and associated online discourse, we identify overlapping and competing understandings of bias, with key fault lines around demographic categories of gender, race, and geography, as well as genres of production and channel size. We analyse how reaction videos navigate structural factors that inhibit discussions of platform practices and assess the functions of aspirational platform governance, including its counter-intuitive role in legitimising content moderation through the airing of complaints.

This paper is part of **Content moderation on digital platforms: beyond states and firms**, a special issue of *Internet Policy Review* guest-edited by Romain Badouard and Anne Bellon.

ACKNOWLEDGEMENTS

The first two authors contributed equally to this work.

Moderation, broadly defined as "governance mechanisms that structure participation" online (Grimmelmann, 2015, p. 47), simultaneously enables cooperation and drives controversy, surfacing enduring disputes over what is acceptable and who gets to decide. While this claim certainly applies to communities formed on the early Web (e.g., Tepper, 1997), the contemporary platform ecosystem introduces new complications. Transnational platforms like YouTube, TikTok, and Meta bring together people with diverse cultural values (Jiang et al., 2021) and pose logistical challenges of cheaply and efficiently moderating millions, even billions, of users. Most platforms respond by outsourcing the work to low-paid contractors (Roberts, 2019) and implementing automated systems (Gorwa et al., 2020), neither of which are well-suited to handle context or nuance. As public life increasingly plays out on privately-owned platforms, content moderation policies and enforcement practices take on greater significance (Van Dijck et al., 2018). For creators who build careers around social media engagement, content moderation also functions as labour management, adding economic considerations to the costs of online exclusion (Are & Briggs, 2023; Ma et al., 2023). Together, these factors reinforce the

conclusion that content moderation is not a game platforms can win, only manage (Gillespie, 2018).

Yet high levels of mistrust reported in surveys (Nicholas, 2022), a spate of user protests (Shapiro et al., 2024; Sybert, 2022), increased scrutiny from the press (Marchal et al., 2024), escalating regulation from the European Union (Keller, 2024), and Silicon Valley's recent "phenomenal capitulation" to the policy priorities of the second Trump administration (DiResta, 2025, 35:35) all point to the "crisis of legitimacy" facing social media platforms (Zuckerman & Rajendra-Nicolucci, 2023, p. 6). Concerns with legitimacy are intimately tied to content moderation, with critics lodging accusations of "bias" and "censorship" toward social media platforms, exploiting the notoriously evasive meaning of both terms (Friedman & Nissenbaum, 1996; Shen et al., 2018) and shifting the target of suppression according to the context. Academic and industry research offers ultimately inconclusive accounts of the presence and direction of platform biases, stymied by the emergent dynamics of complex socio-technical systems and restricted access to relevant data. Amidst this looming legitimacy crisis, the number of active users on mainstream social media platforms continues to grow (Kemp, 2024), raising questions about how users perceive the legitimacy of content moderation and navigate accusations of platform bias.

While content moderation affects all users, its effects are particularly pronounced among creators whose livelihoods depend on social media (Ma et al., 2023). Given their positionality, creators' perspectives offer valuable insights into diverse aspects of content moderation, including the implementation of copyright enforcement (Hallinan et al., 2024), algorithmic governance (Reynolds & Hallinan, 2024), and monetisation programmes (Caplan & Gillespie, 2020). Creator research also surfaces diverse accounts of platform bias as disproportionately affecting Black creators (Harris et al., 2023), sexual content creators (Are & Briggs, 2023; Leybold & Nadegger, 2023), marginalised social media users (Delmonaco et al., 2024), small accounts (Caplan & Gillespie, 2020), and political activists (Riedl et al., 2023), to name but a few. Most studies approach the perception of bias as an individual attribute surfaced through interviews and surveys, or, less frequently, through investigations of ideologically-aligned communities (Cotter, 2024). Such approaches struggle to make sense of competing claims where "participants across the political spectrum" report similar experiences of social media censorship (Haimson et al., 2021, p. 22). Accordingly, there is a need for a broader investigation of the public negotiation of platform bias.

To do so, we turn to a debate about platform bias on YouTube prompted by a

trending video from CoryxKenshin, one of the most successful Black creators on the platform (Gutelle, 2022) with over 21 million subscribers at the start of 2025. We conceptualise the conversation that played out across networks of videos and social media comments as an informal strategy of meta-moderation, or the participatory evaluation of content moderation decisions and policies by users (Lampe & Resnick, 2004). Through a content analysis of 115 videos and associated online discourse, we identify overlapping and competing understandings of bias, with key fault lines around demographic categories of gender, race, and geography, as well as genres of production and channel size. These fault lines reveal that "contested platform governance" applies not only to the user-platform relationship but also to different configurations of users (Sybert, 2022). We analyse how reaction videos, a mechanism of informal meta-moderation, navigate structural factors that inhibit discussions of platform bias. Despite the social and emotional benefits of creator conversations about content moderation, we conclude by arguing that meta-moderation primarily functions as an *aspirational* form of participatory governance wherein creators express the desire to influence content moderation on platforms without any guarantee of success. In doing so, creator conversations often result in legitimating the platform as arbiter of public discourse and cultural production.

From moderating abuse to moderation as abuse

While platforms typically downplay their role in shaping what users see and say on social media, Gillespie convincingly argues that content moderation is "central to what platforms do" (2018, p. 13). Platforms employ moderation "to facilitate cooperation and prevent abuse" in online interactions through mechanisms of exclusion, incentive, organisation, and norm-setting, each of which can involve varied degrees of automation, transparency, and centralisation (Grimmelman, 2015, p. 47). As platforms have scaled up, so too has content moderation, bringing together users, contracted labourers, and algorithmic systems into centralised and opaque arrangements (Gorwa et al., 2020; Roberts, 2019). The responsibility to configure these arrangements falls to a class of trust and safety professionals (Zuckerman & Rajendra-Nicolucci, 2023) who promote industry standards to navigate limited resources and a dynamic regulatory environment (Keller, 2024).

Despite increasing professionalisation, content moderation has been framed as a *source* of abuse, infringing on rights, discriminating against users, perpetuating diverse harms, and badly in need of reform. Contemporary criticisms of moderation cross the political spectrum (Haimson et al., 2021; Nicholas, 2022) and come from multiple sectors of society, including social media users (Ma et al., 2023), civil so-

ciety organisations (e.g., Human Rights Watch, 2023), and government officials (Johnson, 2023). The issue of bias attracts a diverse community of advocates, such as "sex workers[,] conservatives, Black Lives Matter activists, plus-sized influencers, trans folks, and many others" who feel disproportionately affected by platform policies (Nicholas, 2023, pp. 3–4). Recent work on user experiences of content moderation emphasises the personal, social, and economic consequences of being denied access to social media platforms (Are & Briggs, 2023; Ma et al., 2023). Together, these accusations of bias invert the conventional understanding of content moderation as a solution to online abuse, framing it instead as a primary source of abuse.

External assessments of platform bias face practical and theoretical challenges. At the practical level, commercial content moderation is opaque (Crawford & Gillespie, 2016) and creates information asymmetries between platforms, users, and the broader public (Cotter, 2023). These asymmetries are particularly pronounced for moderation techniques that reduce the reach of content or accounts (Gillespie, 2022). While platforms provide some data about content moderation through transparency reports, the utility of these reports has been criticised (Zalnieriute, 2021) and access to standardised, and thus comparable, data is one of the key goals of the European Union's Digital Services Act (Keller, 2024). At the theoretical level, there is disagreement about the meaning of bias and how it should be measured (Friedman & Nissenbaum, 1996). Bias is inherent to any decision-making system, and disaggregating desired bias from objectionable bias is a fundamentally normative matter.

Faced with ubiquitous yet opaque content moderation systems, social media users participate in a "culture of speculative guessing" (Kumar, 2019, p. 8) and develop heuristic understandings influenced by personal experience, cultural discourses, and platform disclosures (Bishop, 2019; Cotter, 2023; Reynolds & Hallinan, 2024). While such heuristics substantively shape platform use (Hallinan & Brubaker, 2021), the theoretical and practical challenges involved in assessing platform bias are even more pronounced among social media users, reflected in divergent claims about the target of bias (Are & Briggs, 2023; Haimson et al., 2021; Nicholas, 2023) and the conspiratorial orientation of platform bias discussions, where ideological suspicion, or "the assumption that platforms are biased against one's personal beliefs" (Riedl et al., 2023, p. 2164), manifests across the political spectrum (see also, Lewis & Christin, 2022; Reynolds & Hallinan, 2024). These accounts suggest that the growing consensus around platform bias contains significant internal contradictions. Yet we know little about how these competing claims play out in practice.

In other words, how do different communities on the same platform negotiate diverse experiences and accounts of platform bias?

Meta-moderation and the collective work of (de)legitimation

We conceptualise "meta-moderation" as the participatory evaluation of content moderation decisions and policies. In so doing, we draw on the term's history, which traces back to community moderation on social news website Slashdot at the end of the 20th century. In the website's parlance, meta-moderation involves "a second layer of moderation" that "seeks to increase fairness by letting logged-in users 'rate the rating' of randomly selected comment posts" (Slashdot, n.d.). The explanation outlines a distributed model of content moderation where users can indicate qualities like "interesting," "overrated," or "offtopic," contributing to the visibility of comments and the reputation scores of commenters. The website incorporates meta-moderation by inviting select users to evaluate decisions as "fair" or "unfair," impacting the reputation scores of moderators and creating an incentive to align with community norms (Lampe & Resnick, 2004). In so doing, meta-moderation formally distributes participation in content moderation in service of creating a "fairer" and thus more legitimate form of governance.

Slashdot's configuration of meta-moderation has received little uptake given its dependence on consensual norms (Lampe & Resnick, 2004). The model fits poorly with contemporary social media platforms which bring together diverse users and typically lack unified reputation scores to incentivise cooperation. Furthermore, because major platforms implement centralised policies developed by trust and safety teams, user ratings of fairness would likely have little effect on policy matters, undermining their legitimating function. Finally, evidence from existing forms of distributed content moderation facilitated through platform reporting tools calls into question the value of flagged data given its potential for manipulation (Crawford & Gillespie, 2016). Despite these challenges, platforms have experimented with alternative configurations of meta-moderation, including Meta and X's brief flirtation with user voting on proposed policy changes (Gillespie, 2018; Race & Miller, 2022), user interactions with the official personification of platform governance on Douyin (He & Tian, 2023), and, perhaps most significantly, the consensus-based evaluation of Community Notes as implemented on X and anticipated on Meta platforms (Kaplan, 2025). However, these approaches are acutely limited and most platforms lack a meaningful "formalised process of stakeholder participation" in decision-making (Kumar, 2019, p. 15).

In the absence of formal channels, social media users employ informal strategies of participation in platform governance. There is an emergent body of research investigating how social media users mobilise publicity to contest content moderation (e.g., Hallinan et al., 2024; Leybold & Nadegger, 2023; Reynolds & Hallinan, 2024; Shapiro et al., 2024; Sybert, 2022), functionally describing meta-moderation without invoking the concept. Informal strategies of meta-moderation are almost entirely deliberative: creators mobilise the communicative affordances of social media for indirect influence, seeking to, though not necessarily succeeding in, setting platform norms (Grimmelman, 2015). Prior work has focused on the mobilisation of individual communities, leaving intra-community dynamics of meta-moderation unexplored. We investigate frictions among users through a public debate about platform bias on YouTube, asking how a diverse group of creators problematised biased content moderation.

Intra-community contestation of content moderation on YouTube

The debate prompted by CoryxKenshin's callout video, which we detail in the next section, acts as a critical case study with "strategic importance in relation to the general problem" of bias in content moderation (Flyvbjerg, 2006, p. 229). In other words, if a creator as prominent as CoryxKenshin, with a track record of good conduct and access to internal contacts within YouTube, *still* struggles to understand the application of the platform's rules, issues with fair and transparent content moderation likely plague other creators. The reactions to Cory's video bear out this hypothesis; his concerns are not an isolated case of misapplied systems but rather reflect structural frustrations around the communication between the platform and creators. While our study therefore stems from CoryxKenshin's discussion of the possible racism and favoritism embedded in YouTube's governance systems, it surfaces similar concerns from creators within and across demographic lines. CoryxKenshin's video ruptured the platform ecosystem where conversations that normally reside behind-the-scenes became, temporarily, front and center.

The case also highlights a rare conversation about the role of race and racism on YouTube led by creators. With a few exceptions, such as the niche communities focused on natural hair care (Sobande, 2020) or Afrofeminism (Da Silva, 2022), YouTube has not produced a kind of Black commons, or "a discursive place that serves as a location for depositing commonplace arguments and persuasive messages that reflect the issues and needs of the Black community" (Steele & Hardy, 2023, p. 317) akin to the collection of users described as Black Twitter (Graham &

Smith, 2016). As a result, many scholarly investigations of racism on YouTube focus on user comments (e.g., Murthy & Sharma, 2019) or media representations of minority creators (e.g., Guo & Harlow, 2014) rather than creator-driven discussions. Yet Black creators have long held concerns about platform biases. For example, YouTube's former Chief Business Officer Robert Kyncl recounted talking to leading Black creators in 2016 about their reluctance to use their faces in video thumbnails, noting that they get fewer clicks and more negative reactions (Kyncl & Peyvan, 2017). However, many of these conversations take place outside of the platform; by contrast, our data set represents a moment of public collective concern about racism and bias that attempts to address platform moderation *on* the platform itself.

YouTube offers an ideal site for investigating creator-driven platform governance because of its emphasis on professional content creation (Hallinan et al., 2024), demonstrated through investments in programmes like YouTube Spaces that provided creators "with important resources, including state-of-the-art studios, events, and classes" (Kyncl, 2021), its early establishment of an appeals process to dispute moderation decisions (Díaz & Hecht-Felella, 2021), and its development of the YouTube Partner Program that provides large creators with internal company contacts known as Partner Managers (Caplan & Gillespie, 2020). YouTube has also expressed commitment to equality on the platform, and to its Black creators in particular, launching the #YouTubeBlack Voices Fund in 2020 that provides resources and support to select Black creators (YouTube Creators, n.d.).[1]YouTube has thus dedicated rhetorical and financial support towards its goal of putting "equity and inclusion at the forefront of our mission" (Mangroo & Paul, 2023). Yet, as our analysis demonstrates, creators have doubts about how YouTube puts these principles to practice.

To map the conversation, we identified relevant videos through targeted keyword searches,[2]references within videos, and automated recommendations. We determined relevance through the video's title, thumbnail, description, or introduction, resulting in a list of 310 videos that we fed to YouTube Data Tools to retrieve

1. Notably, while the creation of the #YouTubeBlack Voices Fund was seemingly spurred by the 2020 reckoning with racial politics in the United States spurred by the murder of George Floyd, the fund does not focus on Black American creators only, but has also issued grants to Black creators from "Kenya, the United Kingdom, Brazil, Australia, South Africa, and Nigeria" (YouTube Creators, n.d.). These locations have distinct racial politics and understandings of what it means to be Black, reflecting YouTube's attempts to foster a global community while still operating as an American company.
2. These included variations on CoryxKenshin's name, the title of the video, the video's URL, and combinations of YouTube, racism, and bias.

metadata (Rieder, 2015). The first author familiarised themselves with the data by watching the introductions of each video. Building on initial observations and previous experience analysing creator callout and audit videos (Hallinan et al., 2024; Reynolds & Hallinan, 2023), they developed a codebook to characterise the identity of creators, attitudes toward claims of racism and bias in content moderation, and types of supporting evidence marshalled (available upon request). We coded the perceived race and gender of anyone who appeared on screen, prioritising self-identification when available (nearly 40% of participants included racial self-identification, compared to only 18% for gender). While relying on conventional visual markers necessarily flattens diverse experiences and expressions of identity, draws on researcher biases, and can constitute epistemic violence (Keyes et al., 2021), identity remains deeply relevant to conversations about race and racism, as well as gender segregation on YouTube (Wegener et al., 2020). We thus use imperfect measures to grapple with the role of identity in the conversation, mitigating harms by only discussing inferred identity categories at the population level and using consensus coding between two authors. The team jointly watched three videos to refine the codebook, after which the first and third authors coded the remainder of the corpus, stopping after reaching theoretical saturation at 115 videos (37% of the total data set). Given the length of some videos, we set a soft limit of watching the first 15 minutes, making exceptions for particularly interesting cases. We then coded the primary genre of each channel based on titles and thumbnails of recent videos.

Calling out content moderation

On 24 August 2022, Cory DeVante Williams, a Black gaming YouTuber better known as CoryxKenshin, uploaded the video "YouTube: Racism and Favoritism," which quickly appeared at the top of the YouTube trending list in the United States (Gutelle, 2022). In the video, CoryxKenshin describes his frustrations with content moderation on the platform as a longtime creator who now has over 21 million subscribers. Despite this remarkable success, CoryxKenshin outlines a history of disparate treatment, exemplified by a recent incident involving his playthrough of *The Mortuary Assistant*. CoryxKenshin was one of many creators who made videos about the newly-released game which positions the player as an employee at a haunted mortuary who must exorcise demons from corpses. Given the macabre subject, it would perhaps be reasonable to expect YouTube to age-restrict playthroughs of the game—yet many were not.[3] Although CoryxKenshin censored curse words from the game and limited his interjections to family-friendly jokes

3. Including, notably, CoryxKenshin's own playthrough of the game's demo version a year prior.

and wordless shouts of surprise, his video was flagged as inappropriate for general audiences per an unspecified portion of YouTube's Community Guidelines, severely reducing its algorithmic reach and monetisation potential. He appealed the decision and a long saga followed, where YouTube first removed the age restriction and then re-restricted the video, and also restricted a video of fellow gaming creator Markiplier after he publicly expressed solidarity with CoryxKenshin.

CoryxKenshin acts as a "perfect plaintiff" within the YouTube governance ecosystem (Godsoe, 2015). Until the incident with *The Mortuary Assistant*, he was a model creator, proactively adhering to content guidelines, building a large audience premised on positive interactions, and never serving as a vector for embarrassing controversy. His public record features a single prior dispute with the platform, reflected in a similar video he uploaded a year earlier calling for better communication between YouTube and creators (see Figure 1). In this regard, he strikingly diverges from other homegrown YouTube stars like PewDiePie and MrBeast, both of whom have been embroiled in scandals that reached mainstream news. CoryxKenshin, by contrast, fulfils what Godsoe (2015) notes as the chief imperative for the public face of controversy: "Be normal" (p. 136). Indeed, the resonance of CoryxKenshin's frustration with YouTube's decision may lie exactly with this reasoning—he behaved normally but was still punished by opaque powers-that-be. This normality makes the direct comparisons CoryxKenshin draws between his treatment and the lack of moderation for Markiplier, an established Asian American creator who works in the same genre, especially striking. However, further escalating perceptions about YouTube's preferential treatment, Markiplier was cut without explanation from YouTube's Game On event, which he had helped to plan and promote (Fisher, 2022), days after he publicly supported CoryxKenshin.

FIGURE 1: Author screenshots of two videos from CoryxKenshin addressing the topic of content moderation.

CoryxKenshin's video played well in the court of human appeal, activating other creators who had felt similarly powerless in the face of opaque content moderation. As Lithwick notes, courts, like most listeners, prefer people who "play by its rules" and tell them "stories they like to hear about people who remind them of themselves" (Lithwick, 2012), two criteria which CoryxKenshin aptly achieves in his callout video. This schema works as long as we consider adjudication a fundamentally human act—which, of course, is often not the case given platform governance's assemblage of human and non-human actors. However, YouTube is not the only adjudicator; CoryxKenshin's audience and peers also serve as judges and pre-

sented a wide range of reactions to his claims of YouTube bias. For these audiences, the title of the video immediately captured attention, implicating YouTube in wrongdoing. The structure continues the suspense, with CoryxKenshin invoking the possibility of racism as a mystery that needs to be squared with a series of suspicious moderation events. Finally, the video hits credible emotional notes—CoryxKenshin is openly frustrated that he feels the need to even make such a video, expresses he did not want to, and describes the many backstage routes he took to address the problem first, including formal appeals and informal conversations with YouTube employees.

Reaction videos as informal organising

CoryxKenshin's video prompted a conversation that extended beyond the comment section—although it also took place there, accumulating almost 100 thousand comments and counting. Other YouTubers, inspired by the video's topic and prominence on the trending list, created and uploaded related videos, broadly engaging in a practice known as "reaction" (McDaniel, 2021). The quintessential reaction video features a creator broadcasting their real-time response to a piece of media, transforming cultural consumption into cultural production. Reacting is a social practice that lacks specific technical support; there is no "react" button that formalises the connection between two pieces of content akin to the "quote" feature on Twitter/X (Burgess & Baym, 2020), the "stitch" feature on TikTok (Literat et al., 2023), or the discontinued "reply" feature on YouTube (Bucher, 2018). Working within the constraints of convention and copyright enforcement, creators integrate media differently, ranging from playing a video in full on the screen to only verbally referencing content—a spectrum fully represented in reactions to CoryxKenshin's video (see Figure 2). Some creators even recursively react to other reaction videos, introducing an additional level of "meta" to the conversation on content moderation.

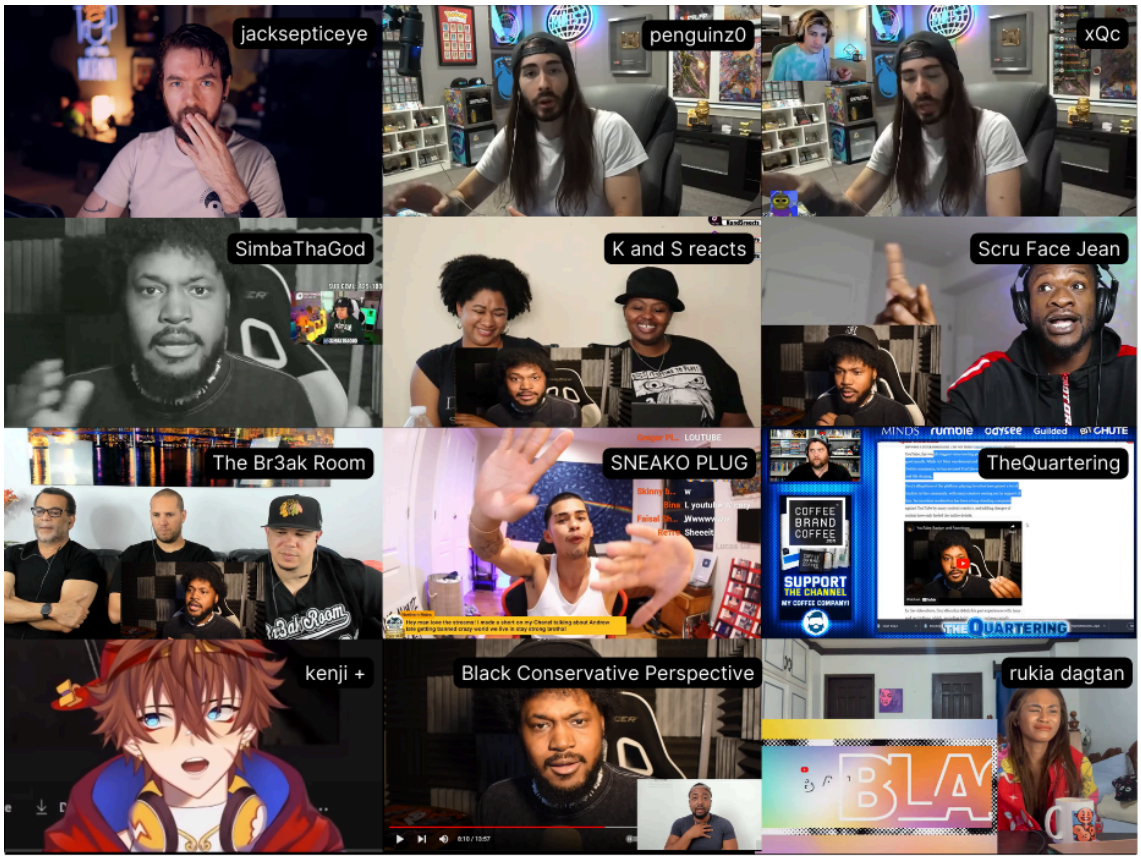

FIGURE 2: A collage of author screenshots illustrating the expressivity, recursivity, and stylistic diversity of reaction videos within the data set.

The network of videos connected through hyperlinks, verbal recommendations, and topics of conversation made tangible a shared concern with content moderation. Yet participation in the conversation was not evenly distributed across YouTube's demographics. Among the videos we analysed, Black men were by far the most prominent, accounting for half of all participants (see Table 1). White men, Black women, and creators with indeterminate identities—including virtual YouTubers and videos without self-representation—were next most prominent; all other categories barely surpassed 5% combined. The prominence of Black creators aligns with the topic of racism, with many invoking their own racial identity to co-sign experiences of racism or, less frequently, push back against adopting, in the words of SNEAKO, a "victim mentality" (2022, 16:38). It also may reflect the influence of reaction as a genre where Black creators have achieved particular success, responding to "a demand for Black affirmation of white viewers' cultural worlds" (Rosen, 2020). The minimal representation of women, particularly non-Black women, aligns with previous studies of platform callout videos (Hallinan et al., 2024) and survey research suggesting that men are more likely to believe that they have been suppressed by social media platforms (Nicholas, 2022). Finally, the geo-

graphic distribution of channels is highly concentrated, with over 63% listing their location as the United States (see Table 2), although creators from around the world may do so in an attempt to reach international audiences (Bidav & Meta, 2024).

TABLE 1: Demographic breakdown of YouTubers based on a combination of disclosed and perceived identities.

IDENTITY	TOTAL	PERCENTAGE
Asian	6	4.8%
Asian men	5	4.0%
Asian women	1	0.8%
Black	76	61.3%
Black men	62	50.0%
Black women	14	11.3%
White	29	23.4%
White men	27	22.8%
White women	2	1.6%
Unclear	13	10.5%

TABLE 2: Number of subscribers, primary genre, and country of sampled YouTube channels.

SUBSCRIBERS		GENRE		COUNTRY	
1-1k	23	Reacting	38	United States	73
1k-10k	20	Commentary	37	Unlisted	22
10k-100k	23	Gaming	30	Other	8
100k-1m	21	Other	10	Canada	6
1m+	27			United Kingdom	6

Beyond conventional demographic categories, creators on YouTube have at least two other prominent sources of community affiliation: genre and channel size. Almost all channels we analysed primarily produced reaction videos, commentary, or videogame content (see Table 2). Given the initial video's trending status and provocative premise, the prevalence of react channels makes sense. Similarly, attention from commentary channels aligns with their respective focus on societal

and platform-based issues (Lewis et al., 2021). Finally, videogame creators emphasised their affinity with CoryxKenshin as a fellow gamer, including several prominent white male gamers. Genre affiliation also shaped the style of reactions (see Figure 2), with react channels likely to re-broadcast the video alongside dramatic facial expressions and commentary channels likely to offer "talking-head" style discussions. Gaming channels were more mixed in their approach, with creators who livestream more likely to adopt the classic react style.

Participants represented a mix of small, medium, and large channels (see Table 2), including 23 very small channels often overlooked in platform research (Da Silva, 2022). YouTube employs a tiered governance strategy that offers "different users different sets of rules, different material resources and opportunities, and different procedural protections" (Caplan & Gillespie, 2020, p. 2). As a successful creator with millions of subscribers, CoryxKenshin is among the creator elite with a designated partner manager who can advise, answer questions, and help resolve disputes. The tiered governance system featured prominently in the networked conversation, with larger creators highlighting the limited influence of partner managers within the company and smaller creators describing the challenges of independently navigating content moderation.

The conversation benefited from the engagement of high-profile individuals from the gaming and commentary communities. Markiplier (36 million subscribers), jacksepticeye (30 million subscribers), and penguinz0 (14.5 million subscribers) were particularly notable supporters, and their four videos on the topic collectively amassed more than 24 million views.[4]Large accounts amplified the conversation and made it more acceptable, even desirable, for others to talk about an otherwise taboo topic. Some creators chose to release their videos on secondary channels to protect the algorithmic recommendations of their primary content while others mentioned that they were only willing to talk about racism because CoryxKenshin's video went trending. The involvement of high-profile individuals and the collective practice of reaction videos were the primary ways of building a conversational network; alternative mechanisms of affiliation such as hashtags were rarely invoked (Graham & Smith, 2016). As a result, the conversation about platform bias and racism remained primarily organised around a specific video.

4. View counts as of 30 January 2025 and include the videos "Big Problems at Youtube" by penguinz0 (https://youtu.be/IPXukSZhTuI?si=VI3b2iRyBcmn6HMx) at 5.8 million views, "Youtube has some serious issues…" by jacksepticeye (https://youtu.be/_IuqFlWovGQ?si=gRWaoRGcRNVH6X8q) at 5.8 million views, "Try Not To Get Age Restricted Challenge" by penguinz0 (https://youtu.be/fByMGb-FUZxE?si=J7plrjvZWB8Jkgvt) at 2.6 million views, and "Try Not To Get AGE-RESTRICTED Challenge" by Markiplier (https://youtu.be/fBJ72GhkuFY?si=3HtJNpvK25z6QwWs) at 10.4 million views.

Problematising platform bias

All reaction videos engaged with CoryxKenshin's callout, amplifying its reach and guiding audience responses, although reactions diverged concerning the video's primary claims about content moderation. Overall, the data suggests that nothing brings people together like feelings of discrimination, with over 80% of participants agreeing that YouTube's content moderation is biased (see Figure 3). However, a closer look at the meaning of bias paints a more complicated picture. According to the discourse, YouTube's biases are based on the following criteria: political ideology (typically favouring "the left"); channel size (typically favouring big channels); the type of content (typically favouring A-list celebrities and legacy media companies while disfavoring dark content, reaction videos, and critical discussions of YouTube); creator demographics (with accusations that the platform both favours and disfavours women and members of the LGBTQ community, as well as more consistent claims of favouring white creators, English speakers, and Americans); and personal relationships with specific creators (including Markiplier, PewDiePie, MrBeast, H3H3, and James Charles). The conceptual openness of "bias" creates space for people with very different views to locate their experiences.

The remaining videos were split between ambivalence and antagonism. Our "can't tell" code includes videos where creators expressed scepticism about the ability to understand YouTube's content moderation and videos where the creator did not take a stance. The latter situation was particularly prominent in classic react videos where creators emphasised affective responses over explicit evaluation. While the facial expressions and body language typically seemed supportive of CoryxKenshin, we refrained from coding implicit signals as either endorsements or refutations. However, the few channels that disagreed with the existence of biased content moderation did so explicitly and vehemently, adopting provocative stances that attracted a lot of attention from CoryxKenshin's legion of fans, known colloquially as The Samurai. For example, SonnyTM's video, titled "@CoryxKenshin: Privileged Crybaby" rejects the narrative of CoryxKenshin's callout, characterising him as "full of shit" to great effect (2022, 0:41), drawing in over 750,000 views for a channel with 11,000 subscribers. Whether supportive, agnostic, or antagonistic, most creators openly expressed their opinions about platform bias. Yet it's notable that women—the overwhelming majority of whom were Black—universally agreed with the existence of bias in content moderation, despite their comparably low rates of participation.

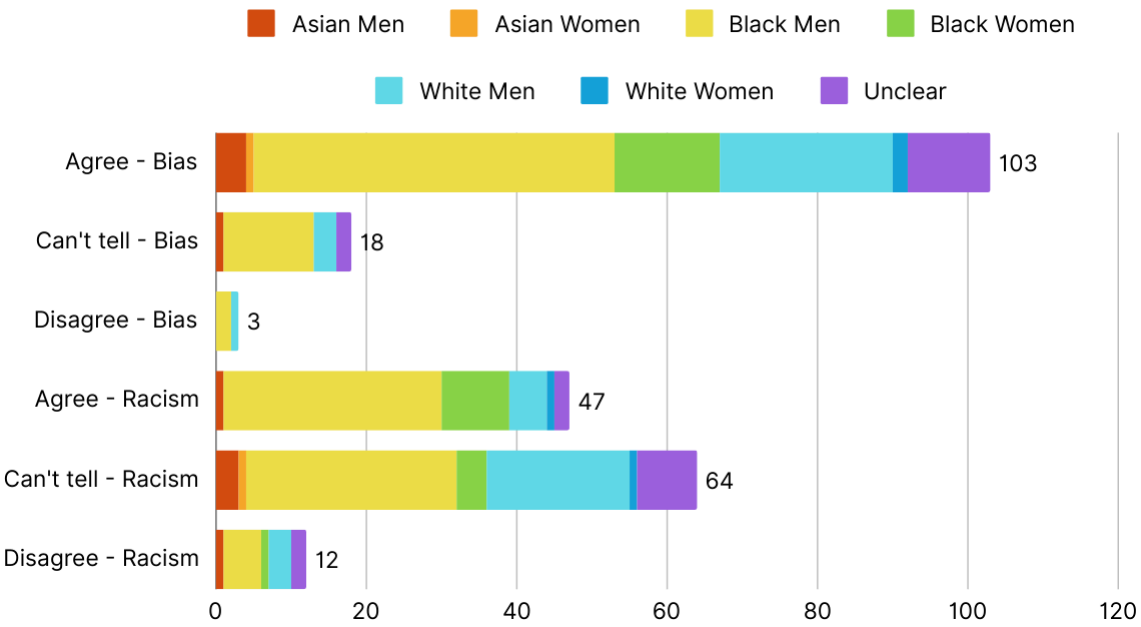

FIGURE 3: Attitudes toward the claims that YouTube's content moderation is (1) biased and/or (2) racist.

The same willingness to weigh in did not apply to discussions of racism. Indeed, the claim that YouTube's content moderation is racist proved far more ambivalent: nearly 38% of participants agreed, 10% disagreed, and the remaining majority adopted an ambivalent stance. Creators often expressed that accusations of platform racism were difficult or even impossible to adjudicate. Miniklin, a creator sympathetic to the claim of racism, still noted that "racism is a huge thing to claim, especially when you don't 100% know" and invited his audience to send proof of platform discrimination (2022, 8:57). Others felt reluctant to speak about racism given their positioning; for example, Rurikhan, a gaming creator from Portugal, explained "I'm not gonna be tackling racism because quite simply I'm a white man. What do I know about racism on the internet?" (2022, 4:04). This concern is reflected in the demographic distribution of support for the existence of racist content moderation, which primarily came from Black creators and a small subset of white men.

The discussion largely follows ways of speaking about race and racism typical to the United States, including an emphasis on diverging Black and white experiences (Yancy, 2017). These patterns are evidenced in the ambiguity surrounding Markiplier as an Asian American: some videos frame him as a person of colour and cite the platform's willingness to age-restrict his video as further evidence of racism while others describe him as white and use that identity to explain his favoured treatment. Overall, there is little discussion of intersectionality (Da Silva,

2022), or a consideration of how different identities form interlocking systems of oppression. However, there were a few exceptions to this rule. For example, Asamara explained that as a "Black woman, you see the differences. This world was set up for white success, not just on YouTube but in general in the workforce whether or not you're part of the creator economy" and criticised YouTube's Black Voices Fund as an attempt at "diversity without inclusion" (2022, 3:41). Similarly, rukia dagtan, a prominent react creator from the Philippines, emphasised the importance of geography, expressing frustration with how YouTube favours American creators (2022).

Limited access to information about content moderation simultaneously limits creators' ability to make convincing claims about structural conditions on the platform. CoryxKenshin openly admits the difficulty of knowing with certainty whether (or how) race factors into his experiences of content moderation, and the preponderance of creators who concluded that it was impossible to know whether the platform "is racist" emphasised the epistemological challenges involved. As gamer and teacher Azkhalon explains, contemporary racism is "very disguised nowadays, it's hidden, like you know you don't know the real reason why CoryxKenshin's video was age restricted and you'll never know" (2022, 18:02). The most prominent sources of evidence invoked in the videos were personal testimony (n=70) and comparisons to other creators (n=61). Indeed, many of the people who agreed with the assessment of platformed racism also reported personal experiences with racial discrimination. A similar emphasis on personal experiences played out in discussions of bias. These patterns of evaluation demonstrate the limitations of using personal testimony, even networked testimonies, to document and intervene in structural experiences on the platform. Creators from different backgrounds broadly share a desire for "equal" treatment on the platform but disagree quite radically about what that means and what evidence of its violation looks like.

Informal meta-moderation as aspirational platform governance

The public debate surrounding CoryxKenshin's allegations of platform bias was a clear attempt to evaluate the fairness of content moderation on YouTube, analogous in spirit to the model of meta-moderation developed in an earlier internet era. Yet the informal, creator-driven discussion enabled by vernacular practices like reaction videos bears little resemblance to formalised community governance on Slashdot. The differences are both structural and cultural. Structurally, user evaluations of content moderation decisions on Slashdot directly impacted the reputa-

tion score of the moderator while public discussions on YouTube attempt to marshall indirect influence through publicity and moral appeal (Shapiro et al., 2024). Culturally, participants in Slashdot's meta-moderation programme demonstrated remarkably strong community consensus (Lampe & Resnick, 2004); the same cannot be said for YouTubers' concerns with content moderation. While we found significant support for the idea that content moderation is biased, creators substantively disagreed on the targets and mechanisms of biased treatment—an unsurprising situation given the size and diversity of YouTube as a platform, even as our case study focuses on a conversation taking place exclusive among English-speaking creators primarily making gaming, reaction, and commentary content. However, both Slashdot's formalised system of meta-moderation and YouTube's culture of informal participation through content creation negotiate a desire on the part of users to engage with platform decision-making.

The manifestation of meta-moderation on YouTube primarily functions as aspirational platform governance, expressing the desire to influence content moderation policies and practices without established channels or guarantees of success. In publicly challenging the platform's content moderation decisions, YouTube creators refuse to take "for granted the definition of the problem and the aims of the stakeholders" (Gillespie, 2023). This is a form of digital civic engagement, acted out through creators' use of their position of value to the platform and their ability to marshall the attention of pre-developed audiences. In terming these practices "aspirational platform governance," we bring together Kligler-Vilenchik and Literat's (2024) account of expressive citizenship and Duffy's (2017) formulation of aspirational labour. According to Kligler-Vilenchik and Literat, social media platforms have transformed conventional modes of civic engagement into a "new, emergent citizenship model" (2025, p. 47) in which "publicly expressing views around current events is a form of political participation in and of itself" and "normatively expected of 'good citizens'" (2024, p. 131). This imperative to weigh in on current events may be even more marked for creators, as public communication is intertwined with their labour and livelihoods.

Shifting attention to entrepreneurial activity, Duffy argues that creators engage in "(mostly) uncompensated, independent work that is propelled by the much-venerated ideal of getting paid to do what you love" (2017, p. 4), highlighting the cultural importance of a belief in future rewards. The participatory promise of meta-moderation is influence over the policies and practices of platforms into which creators have invested significant time and labour. However, lacking formal mechanisms of participation, and often having first exhausted more private forms of

protest like appeals, public discussions of governance remain predominantly aspirational, hopeful that enough noise will catch the right ear within platform leadership to provoke change. As YouTuber jacksepticeye optimistically puts it, “The whole internet seems to have been rallying against YouTube on their own platform in the last week… I think that that's how change gets made because otherwise it just gets lost and then no one ever talks about it” (2022, 0:15).

Informal strategies of meta-moderation are also aspirational in the sense that there is minimal evidence this approach directly changes platform policies. YouTube did not publicly acknowledge CoryxKenshin's video or the ensuing conversation among creators, although it has occasionally acknowledged previous policy concerns raised by creators (Hallinan et al., 2024; Reynolds & Hallinan, 2024; Shapiro et al., 2024). Given the challenges of tracing influence within large organisations, especially from the outside, it is worth acknowledging the ambiguous effectiveness of aspirational governance. Even if singular incidents do not provoke policy change, a series of incidents may guide corporate strategy. For example, the repeated public concern with poor communication about content moderation decisions may well have influenced two major changes YouTube recently implemented: the addition of timestamps for Community Guidelines violations, and a guided resolutions flow designed to clarify and streamline the appeals process (Creator Insider, 2023). While the weight of any given outcry is difficult to measure, there is some evidence that the repetition of collective concerns over time may influence policy changes (see also, Marchal et al., 2024).

However, even if aspirational governance does not affect YouTube's decisions, such informal organising still provides social and emotional support for creators, affirming that the difficulties they face on the platform are not solely their own (Leybold & Nadegger, 2022; Ma et al., 2023). Bringing content moderation experiences into the open and prompting collective discussion is particularly valuable given the individualising structure of YouTube, where ideas of community are often about organising audiences *around* particular creators rather than organising *among* creators. This may help to explain why our data set contained unusual parity between large and small channels, as publicly sharing experiences offers a partial response to the information inequalities built into commercial content moderation, potentially resisting what Cotter calls “black box gaslighting” (2023) and countering feelings of loneliness and isolation tied to entrepreneurial labour. Further, the sharing of channel analytics and personal experiences may help to create cracks in the structural asymmetries of power and information between creators and platforms (Cutolo & Kenney, 2021). YouTubers regularly point to poor communication

from the platform regarding content moderation and policy decisions as primary concerns surrounding their labor and content choices (Reynolds & Hallinan, 2024). Reaction content centered around a topic of collective concern, like the possibility of racism or favoritism in platform decision making, contributes to information exchange that mitigates creators' uncertainty about whether the concerns they have with content moderation are isolated or widespread. Thus, while creators in our data set readily concede that they can never truly know whether a given decision was motivated by some form of bias, collectively their reaction videos serve to share experiences, trace patterns of enforcement, and strategise responses to common problems. In so doing, they both seek to place pressure on content moderation decisions and develop community norms, while also hoping that these efforts will contribute to structural changes to platform governance.

Ultimately, we argue that aspirational platform governance legitimates companies like YouTube, even though callouts traffic in public criticism. Creators routinely expressed their belief in YouTube's right to make moderation decisions and noted that many people within the company are genuinely trying to do their best for creators. As such, criticisms instead focused on the belief that YouTube can and should be better at content moderation, with improvements driven by listening to creator experiences and enabling their participation. This is not an attempt to rebel against the existence of rules writ large, but to shape specific rules and their enforcement. In calling for these changes, some creators positioned themselves more antagonistically while others, like CoryxKenshin, adopted a more collaborative appeal. The caption to CoryxKenshin's 2021 video, framing it as a "conversation" between platform and creator, exemplifies this attitude—and in a conversation, willing listeners and fair responses are desired, even as creators qualify the extent to which they actually expect the platform to respond. As such, public critiques of YouTube's content moderation decisions are fundamentally different from attempts to subvert the platform or challenge its right to moderate; the aspiration of participating creators is instead to be formally recognised as an important part of the system.

Conclusion

Overall, our study offers the first analysis of reaction videos as a mechanism of informal organising and contributes to an understanding of how users navigate the "crisis of legitimacy" (Zuckerman & Rajendra-Nicolucci, 2023, p. 6) facing social media platforms. Theoretically, we rework the concept of meta-moderation to make sense of emergent and informal practices of participation in centralised con-

tent moderation, yet caution against overly optimistic assessments of its influence by characterising meta-moderation as a form of aspirational platform governance. Empirically, we map the strategies and concerns expressed by a diverse group of creators in a rare public discussion about the role of race and racism on YouTube. Methodologically, we outline an approach for tracking informal organising on social media beyond the hashtag. Yet our account of the problem of bias in content moderation remains limited by our focus on the public nature of aspirational platform governance. Discussions of racism and other biases are still somewhat taboo among creators who must concern themselves with appealing to sponsors, audiences, and recommendation algorithms. There are also persistent gaps in gendered participation in platform callouts, with both this study and our previous research demonstrating the overall rarity of women choosing to "speak up" regarding platform governance concerns (Hallinan et al., 2024; Reynolds & Hallinan, 2024). Finally, while our analysis focussed on YouTube, aspirational governance is not limited to one platform. Twitch, for example, not only has a long history of creator outcry about moderation decisions, but policy decisions can often be directly traced to these public conversations (Shapiro et al., 2024). And, as noted previously, both X and Meta have experimented with the idea of user votes on policy proposals and are increasingly promoting the Community Notes mechanism of norm-setting. Indeed, the participatory structure of content creation platforms enables and even encourages meta-moderation discussions. Nevertheless, few platforms have formalised public routes of participation in content moderation, choosing instead to invest in signalling tools like flagging (Crawford & Gillespie, 2016) that privatise such concerns by delegating the decision to report to individual users and choosing how to act on reports behind-the-scenes. Our case thus demonstrates what issues may prompt users to abandon platforms' preferred private routes of challenging moderation and instead aspire to publicly participate in platform governance.

References

Are, C., & Briggs, P. (2023). The emotional and financial impact of de-platforming on creators at the margins. *Social Media + Society*, *9*(1). https://doi.org/10.1177/20563051231155103

Asmara (Director). (2022, August 30). *CoryxKenshin called out YouTube (they want diversity without inclusion) | a black girl's response* [YouTube video]. https://youtu.be/7G3SLBAkHUU?si=pRWlzomQo8IXuYHU

Azkhalon (Director). (2022, August 31). *My thoughts .... | Reacting to CoryxKenshin's YouTube: Racism and favoritism video* [YouTube video]. https://youtu.be/71PUar9OET8?si=47G27lO6bxGUxs88

Bidav, T., & Mehta, S. (2024). Peripheral creator cultures in India, Ireland, and Turkey. *Social Media + Society*, *10*(1). https://doi.org/10.1177/20563051241234693

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. *New Media & Society*, *21*(11–12), 2589–2606. https://doi.org/10.1177/1461444819854731

Bucher, T. (2018). Cleavage control: Stories of algorithmic culture and power in the case of the YouTube 'Reply Girls'. In Z. Papacharissi (Ed.), *A networked self and platforms, stories, connections* (pp. 125–143). Routledge.

Burgess, J., & Baym, N. K. (2020). *Twitter: A biography*. New York University Press.

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, *6*(2). https://doi.org/10.1177/2056305120936636

CoryxKenshin (Director). (2022, August 24). *YouTube: Racism and favoritism* [YouTube video]. https://www.youtube.com/watch?v=GaHcnPDcUOE

Cotter, K. (2023). "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, *26*(6), 1226–1243. https://doi.org/10.1080/1369118X.2021.1994624

Cotter, K. (2024). Practical knowledge of algorithms: The case of BreadTube. *New Media & Society*, *26*(4), 2131–2150. https://doi.org/10.1177/14614448221081802

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Creator Insider (Director). (2023, January 20). *Now launching: Timestamps for community guidelines and guided resolution flow!* [YouTube video]. https://youtu.be/91GpKThpVVw?si=ZVa4d7S0m7KWRICu

Cutolo, D., & Kenney, M. (2021). Platform-dependent entrepreneurs: Power asymmetries, risks, and strategies in the platform economy. *Academy of Management Perspectives*, *35*(4), 584–605. https://doi.org/10.5465/amp.2019.0103

Da Silva, J. (2022). Um conceito na rede: Interseccionalidade e sua tradução micromidiática na web francesa [A concept on the web: The translation of intersectionality on the French web]. *Revista Fronteiras, Estudos Midiáticos*, *24*(1), 22–36.

Delmonaco, D., Mayworm, S., Thach, H., Guberman, J., Augusta, A., & Haimson, O. L. (2024). 'What are you doing, TikTok?': How marginalized social media users perceive, theorize, and 'prove' shadowbanning. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–39. https://doi.org/10.1145/3637431

Díaz, Á., & Hecht-Felella, L. (2021). *Double standards in social media content moderation*. Brennan Center for Justice at New York University School of Law. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

DiResta, R. (n.d.). *The "working the refs" edition* [Audio podcast episode]. Lawfare Media. https://www.lawfaremedia.org/article/rational-security--the--working-the-refs--edition

Duffy, B. E. (2017). *(Not) getting paid to do what you love: Gender, social media, and aspirational work*. Yale University Press.

Fisher, C. (2022, August 28). *Markiplier throws shade at YouTube after segment was cut from Game On event*. https://www.dexerto.com/entertainment/markiplier-throws-shade-at-youtube-after-segment-was-cut-from-game-on-event-1916515/

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, *12*(2), 219–245. https://doi.org/10.1177/1077800405284363

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2022). Do not recommend? Reduction as a Form of content moderation. *Social Media + Society*, *8*(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

Gillespie, T. (2023). The fact of content moderation; or, let's not solve the platforms' problems for them. *Media and Communication*, *11*(2). https://doi.org/10.17645/mac.v11i2.6610

Godsoe, C. (2015). Perfect plaintiffs. *Yale Law Journal Forum*, *136*. http://www.yalelawjournal.org/forum/perfect-plaintiffs

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Graham, R., & Smith, 'Shawn. (2016). The content of our #Characters: Black Twitter as counterpublic. *Sociology of Race and Ethnicity*, *2*(4), 433–449. https://doi.org/10.1177/2332649216639067

Grimmelmann, J. (2017). *The virtues of moderation*. LawArXiv. https://doi.org/10.31228/osf.io/qwxf5

Guo, L., & Harlow, S. (2014). User-generated racism: An analysis of stereotypes of African Americans, Latinos, and Asians in YouTube videos. *Howard Journal of Communications*, *25*(3), 281–302. https://doi.org/10.1080/10646175.2014.925413

Gutelle, S. (2022, August 25). *Did CoryxKenshin catch YouTube's content moderation team playing favorites?* Tubefilter. https://www.tubefilter.com/2022/08/25/cory-x-kenshin-youtube-content-moderation-racism-favoritism-markiplier/

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–35. https://doi.org/10.1145/3479610

Hallinan, B., & Brubaker, J. R. (2021). Living with everyday evaluations on social media platforms. *International Journal of Communication*, *15*, 1551–1569.

Hallinan, B., Reynolds, C., & Rothenstein, O. (2024). Copyright callouts and the promise of creator-driven platform governance. *Internet Policy Review*, *13*(2). https://doi.org/10.14763/2024.2.1770

Harris, C., Johnson, A. G., Palmer, S., Yang, D., & 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. https://doi.org/10.1145/3610169

He, R., & Tian, H. (2023). Social media influencer and source credibility: Endorsing content moderation on Douyin. *International Journal of Communication*, *17*, 5760–5780.

Human Rights Watch. (2023). *Meta's broken promises: Systematic censorship of Palestine content on Instagram and Facebook*. Human Rights Watch. https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and

Jacksepticeye (Director). (2022, August 31). *The Youtube issues get worse* [YouTube video]. https://youtu.be/JbkdvIauRYM?si=fhnyu66nXam9sabd

Jiang, J. A., Scheuerman, M. K., Fiesler, C., & Brubaker, J. R. (2021). Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, *16*(8), e0256762. https://doi.org/10.1371/journal.pone.0256762

Johnson, A. (2023, October 26). *The facts behind allegations of political bias on social media*. ITIF: Information Technology and Innovation Foundation. https://itif.org/publications/2023/10/26/the-facts-behind-allegations-of-political-bias-on-social-media/

Kaplan, J. (2025, January 7). *More speech and fewer mistakes*. Meta newsroom. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Keller, D. (2024, October 16). *The rise of the compliant speech platform*. Lawfare. https://www.lawfaremedia.org/article/the-rise-of-the-compliant-speech-platform

Kemp, S. (2024, January 31). *Digital 2024: 5 billion social media users*. We are social. https://wearesocial.com/uk/blog/2024/01/digital-2024-5-billion-social-media-users/

Keyes, O., May, C., & Carrell, A. (2021). You keep using that word: Ways of thinking about gender in computing research. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–23. https://doi.org/10.1145/3449113

Kligler-Vilenchik, N., & Literat, I. (2024). *Not your parents' politics: Understanding young people's political expression on social media*. Oxford University Press.

Kligler-Vilenchik, N., & Literat, I. (2025). Expressive citizenship: Youth, social media, and democracy. *Journal of Children and Media*, *19*(1), 46–52. https://doi.org/10.1080/17482798.2024.2438680

Kumar, S. (2019). The algorithmic dance: YouTube's adpocalypse and the gatekeeping of cultural content on digital platforms. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1417

Kyncl, R. (2021, February 18). Reaching more creators and artists through YouTube Spaces. *YouTube Official Blog*. https://blog.youtube/news-and-events/youtubespaces-update/

Kyncl, R., & Peyvan, M. (2017). *Streampunks: YouTube and the rebels remaking media*. HarperBusiness.

Lampe, C., & Resnick, P. (2004). Slash(dot) and burn: Distributed moderation in a large online conversation space. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 543–550. https://doi.org/10.1145/985692.985761

Lewis, R., & Christin, A. (2022). Platform drama: "Cancel culture," celebrity, and the struggle for accountability on YouTube. *New Media & Society*, *24*(7), 1632–1656. https://doi.org/10.1177/14614448221099235

Lewis, R., Marwick, A. E., & Partin, W. C. (2021). "We dissect stupidity and respond to it": Response videos and networked harassment on YouTube. *American Behavioral Scientist*, *65*(5), 735–756. https://doi.org/10.1177/0002764221989781

Leybold, M., & Nadegger, M. (2024). Overcoming communicative separation for stigma reconstruction: How pole dancers fight content moderation on Instagram. *Organization*, *31*(6), 879–906. https://doi.org/10.1177/13505084221145635

Linke, C., Prommer, E., & Wegener, C. (2020). Gender representations on YouTube: The exclusion of female diversity. *M/C Journal*, *23*(6). https://doi.org/10.5204/mcj.2728

Literat, I., Boxman-Shabtai, L., & Kligler-Vilenchik, N. (2023). Protesting the protest paradigm: TikTok as a space for media criticism. *The International Journal of Press/Politics*, *28*(2), 362–383. https://doi.org/10.1177/19401612221117481

Lithwick, D. (2012, March 4). Extreme makeover. *The New Yorker*. http://www.newyorker.com/magazine/2012/03/12/extreme-makeover-dahlia-lithwick

Ma, R., You, Y., Gui, X., & Kou, Y. (2023). How do users experience moderation?: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–30. https://doi.org/10.1145/3610069

Mangroo, N., & Paul, S. (2023, February 27). Building equity into our products and policies through the Inclusion working group. *YouTube Official Blog*. https://blog.youtube/inside-youtube/building-equity-into-our-products-and-policies-through-the-inclusion-working-group/

Marchal, N., Hoes, E., Klüser, K. J., Hamborg, F., Alizadeh, M., Kubli, M., & Katzenbach, C. (2025). How negative media coverage impacts platform governance: Evidence from Facebook, Twitter, and YouTube. *Political Communication*, *42*(2), 215–233. https://doi.org/10.1080/10584609.2024.2377992

McDaniel, B. (2021). Popular music reaction videos: Reactivity, creator labor, and the performance of listening online. *New Media & Society*, *23*(6), 1624–1641. https://doi.org/10.1177/1461444820918549

Miniklin (Director). (2022, August 27). *Coryxkenshin just exposed the shocking truth about YouTube..(WTF)* [YouTube video]. https://www.youtube.com/watch?v=h_V9p_XQjoM

Murthy, D., & Sharma, S. (2019). Visualizing YouTube's comment space: Online hostility as a networked phenomena. *New Media & Society*, *21*(1), 191–213. https://doi.org/10.1177/1461444818792393

Nicholas, G. (2022). *Shedding light on shadowbanning*. Center for Democracy & Technology. https://cdt.org/insights/shedding-light-on-shadowbanning/

Nicholas, G. (2023). *Sunsetting "shadowbanning"* (ISP-WIII Essay Series, pp. 1–11). Yale Law School. https://law.yale.edu/isp/publications/platform-governance-terminologies

Race, M., & Miller, M. (2022, December 21). *Elon Musk: Only blue tick users to vote in Twitter polls on policy*. BBC. https://www.bbc.com/news/business-64034892

Reynolds, C., & Hallinan, B. (2024). User-generated accountability: Public participation in algorithmic governance on YouTube. *New Media & Society*, *26*(9), 5107–5129. https://doi.org/10.1177/14614448241251791

Rieder, B. (2015). *YouTube data tools*. YouTube Data Tools. https://tools.digitalmethods.net/netvizz/youtube/

Riedl, M. J., Martin, Z. C., & Woolley, S. C. (2024). 'I get suppressed:' pro- and anti-abortion activists' folk theories of platform governance and shadowbanning. *Information, Communication & Society*, *27*(11), 2153–2170. https://doi.org/10.1080/1369118X.2023.2289976

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Rosen, J. (2020, August 27). The racial anxiety lurking behind reaction videos. *The New York Times Magazine*. https://www.nytimes.com/2020/08/27/magazine/the-racial-anxiety-lurking-behind-reaction-videos.html

rukia dagtan (Director). (2022, August 27). *Reacting to Coryxkenshin YouTube: Racism and favoritism* [YouTube video]. https://youtu.be/YlxBGlmH22c?si=nqMgd9HzIYvORBH5

Rurikhan (Director). (2022, August 29). *Massive problem at YouTube, creators speak out* [YouTube video]. https://youtu.be/yh5eoYPq5Eg?si=cQnakbWhprGavkN6

Shapiro, A., Pippert, C., Smith, J. K., & Taylor, Z. A. (2024). Patrons of commerce: Asymmetrical reciprocity and moral economies of platform power. *Information, Communication & Society*, *27*(10), 1884–1905. https://doi.org/10.1080/1369118X.2024.2331753

Shen, Q., Yoder, M., Jo, Y., & Rose, C. (2018). Perceptions of censorship and moderation bias in political debate forums. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://doi.org/10.1609/icwsm.v12i1.15002

Slashdot. (n.d.). *Moderation and metamoderation. Frequently asked questions*. Slashdot. https://slashdot.org/faq/metamod.shtml

Sneako Plug (Director). (2022, August 24). *Sneako reacs to Corykenshin* [YouTube video]. https://youtu.be/eBzQg0hi-Dw?si=FXIcvUIK5P9DnlCO

Sobande, F. (2020). Black women's digital diaspora, collectivity, and resistance. In F. Sobande, *The digital lives of black women in Britain* (pp. 101–129). Springer International Publishing. https://doi.org/10.1007/978-3-030-46679-4_4

SonnyTM (Director). (2022, September 5). *@CoryxKenshin: Privileged crybaby. YouTube* [YouTube video]. https://youtu.be/SSswXuMymL8?si=NgsSL-o0-UCO0GVc

Steele, C. K., & Hardy, A. (2023). "I wish I could give you this feeling": Black digital commons and the rhetoric of "the corner". *Rhetoric Society Quarterly*, *53*(3), 316–327. https://doi.org/10.1080/02773945.2023.2200704

Sybert, J. (2022). The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *New Media & Society*, *24*(10), 2311–2331. https://doi.org/10.1177/1461444821996715

Tepper, M. (1997). Usenet communities and the cultural politics of information. In D. Porter (Ed.), *Internet culture* (1st ed., pp. 39–54). Routledge. https://doi.org/10.4324/9780203699560

Van Dijck, J., Poell, T., & Waal, M. de. (2018). *The platform society*. Oxford University Press.

Yancy, G. (2017). *Black bodies, white gazes: The continuing significance of race in America* (2nd ed.). Rowman & Littlefield.

YouTube. (n.d.). *#YouTube black voices fund*. YouTube creators. https://www.youtube.com/creators/youtubeblack/

Zalnieriute, M. (2021). Transparency washing" in the digital age: A corporate agenda of procedural fetishism. *Critical Analysis of Law*, *8*(1), 39–53. https://doi.org/10.33137/cal.v8i1.36284

Zuckerman, E., & Rajendra-Nicolucci, C. (2023). From community governance to customer service and back again: Re-examining pre-web models of online governance to address platforms' crisis of

legitimacy. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231196864

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

cnrs centre internet et société

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies