Platform Gaslighting: A User-Centric Insight into Social Media Corporate Communications of Content Moderation

Abstract

This paper delves into communications dynamics between social media platforms and users as they negotiate the complexities of governance policies. Using Meta and TikTok as case studies, we reveal that *gaslighting* - traditionally associated with relationship abuse where one partner undermines the validity of the other's experience - is a pervasive platforms' communications strategy, manifesting in numerous instances where *automated* and *human* platform communications have directly contradicted users' experiences, evidence, and research. We analyse 36 diverse interview datasets and six public platform responses to governance issues, highlighting the systemic nature of this phenomenon within digital spaces. We therefore broaden the scholarly understanding of platform gaslighting by delving beyond shadowbanning and the isolated platform-to-user dialogue to explore a wider range of communications concerning governance. Our participants' experiences show that *gaslighting* can be used to highlight corporate power imbalances in platform-user interactions, especially in situations of opaque governance following not just shadowbanning, but also de-platforming on the back of malicious flagging. Our dataset draws from seemingly disparate groups who share moderation experiences: Jewish creators engaged in combating antisemitism, Palestinian creators advocating for human rights, and sex-positive creators, whose expertise and stories are dismissed and belittled by platforms as a form of damage control in the face of adverse governance. We demonstrate how the dismissal or minimization of participants' traumatic experiences by platforms' automated processes and human teams is weaponized to inflict epistemic injustice, consolidate power, and evade accountability.

Keywords: platform governance, gaslighting, algorithms, shadowbanning, censorship

Introduction

Social media platforms and their users often engage in complex interactions, especially during the actioning and communicating of governance procedures. Platform governance is essential for maintaining social media safety and usability, yet platform companies often receive criticism for content moderation errors leading to unfair censorship (Diaz and Hecht-Felella, 2021), and for the inadequate communication and opacity surrounding their moderation decisions, leaving users feeling unheard (Are & Briggs, 2023; Cotter, 2023). In this paper, we scrutinise Instagram and TikTok's communication strategies, focusing on how they address issues arising from their governance with users and the public. We analyse users' perceptions of platforms' responses to online harms and censorship, showing they perceive Instagram and TikTok's automated and human responses to moderation as a form of gaslighting.

Gaslighting is a term originally coined to describe a form of relationship abuse where one partner systematically undermines the other's beliefs through denial, minimization, or ridicule (Spear, 2019). It features a power imbalance where one partner belittles, minimises, and ridicules the other's personal experiences to the point where the target believes they imagined any wrongdoing. The term has become a staple in dating and general vernacular, but it is now used beyond intimate relationships, extending to fields such as corporate communications. A prime example is the United Kingdom Post Office Scandal, where hundreds of postmasters were wrongfully accused of fraud due to the company's faulty software. The Post Office denied any wrongdoing, falsely attributing any fraud to isolated postmaster mistakes and convicted several innocent workers, some of whom took their own lives as a result (Growns et al., 2024).

As a term rooted in power imbalances, gaslighting is highly relevant to platform governance. Whether deployed intentionally or unintentionally, a communication strategy that uses denial to address complaints can diminish the complainant, leading them to believe they imagined the situation, misunderstood it, or that it resulted from their own actions, and therefore requires scrutiny.

Activists and scholars (Blunt et al., 2020; Cotter, 2023) have adopted this term to depict power imbalances in platform-user relationships, particularly in cases of algorithmic demotion (e.g., shadowbanning) that are dismissed or denied by the platform despite mounting evidence, causing psychological distress to those affected (Are, 2021; Are & Briggs, 2023).

In this paper, we expand on the concept of platform gaslighting, defining it as a user-perceived systematic discursive strategy embedded within platforms' automated and human communication of content moderation processes with users, extending far beyond isolated instances. In doing so, we extend the field-defining scholarship on platform governance by expanding the concept of gaslighting to encompass more than shadowbanning (Blunt et al., 2023; Cotter, 2023), incorporating a wider range of platforms actions such as de-platforming, flagging, and both automated and human communication processes. We argue that opaque communication becomes a stand-in for genuine transparency, fostering a sense of gaslighting among users, even in the absence of overtly malicious intent from the platforms themselves.

Our paper draws upon interview data with three diverse groups of creators: (i) Jewish creators seeking to counter antisemitism and improve religious understanding; (ii) Palestinian creators advocating for peace and resisting discrimination and (iii) content creators posting content related to sex, reproductive and mental health, trans rights and erotic art. The interviews feature TikTok's handling of antisemitism reported by Jewish creators, their shadowbanning of Palestinian creators and censorship of sexual expression, and Instagram's responses to accusations that malicious flagging triggers the de-platforming of nude, sex-positive, and activist creators.

Although these case studies might seem disparate, we group them together due to the similarities in platforms' communication strategies, revealing two discernible patterns of *automated* and *human* gaslighting during platform interactions with users. These patterns include, among others, the denial of specific governance issues, the consistent undermining or discrediting of individual and communal user experiences, a notable lack of accountability, and insufficient communication support for individual creators.

We supplemented our data with a thematic analysis of six key public statements made by platforms to the media: TikTok's reaction to accusations of antisemitism and bias towards the Palestinian perspective following the events of October 7, escalating to Israel's war on Gaza; Meta's default categorization of Palestine-related content as terror in October 2023; Meta Communications' response to users' complaints about governance of sex work following the introduction of new terms of use in December 2020; and the platform's response to widespread account deletions within London's kink, queer, sex-positive, and sex-working communities in June 2023. By drawing attention to public responses, we demonstrate how strategy identification in reaction to opaque practices allows users to collectively recognize these patterns in isolated instances of gaslighting, exposing broader corporate strategies. This approach not only uncovers hidden mechanisms of platform governance but also amplifies user voices, holding platforms accountable for their actions.

Our findings demonstrate that the strategic opacity embedded in both *automated* and *human* communications between platforms and users extends beyond individual interactions to encompass Instagram and TikTok's broader engagements with the media and public. This lack of transparency—frequently framed as a safeguard for corporate interests (Gillespie, 2022)—reflects a pervasive pattern within these companies' communication strategies, prioritising business confidentiality over accountability. Thus, drawing from these diverse instances, we conclude that platforms' lack of transparency and the discursive strategies used to communicate governance decisions serve as a strategic shield to maintain monopolies, fend off public scrutiny, safeguard companies' public image, and sidestep accountability. Their deliberate use of language, narratives, and dialogue maintains pervasive opacity, inflicting epistemic injustice on users by preventing them from fully understanding or challenging platforms' actions, while simultaneously undermining the transparency and fairness of platform governance.

Platforms, Moderation and Harms

Together with reputational damage control and earnings protection, social media harm reduction is one of the main drivers of platform governance (Are & Briggs, 2023). Our paper centres on sociotechnical harm, which "inflicts psychological damage towards a person or community and compromises their ability to participate safely and equitably both online and offline" (Schoenebeck and Blackwell, 2021: 24). Sociotechnical harm emerges at the intersection of gender, race, and disability, revealing how platform development and user experience are architecturally intertwined with cultural forces and systemic inequalities (Shelby et al., 2021). Thus, social media platforms facilitate various harms, including sexual harassment, hate speech, racism, and disinformation (Wood, 2021), that can be intentional (e.g., doxxing) or unintentional (e.g., inaccessible content) and are often exacerbated by societal norms and power structures embedded in technology.

Schoenebeck and Blackwell (2021) identified two primary, overlapping categories of harms: (1) platform-perpetrated harms, resulting from platform design and processes, and (2) platform-enabled harms, which are executed by users or groups but facilitated by platforms. An example is Instagram's failure to act on 9 out of 10 users reported violent threats received via direct messages (DMs), and failure to respond to image-based sexual abuse reports within the first 48 hours (Centre for Countering Digital Hate, 2022).

The labour of gatekeeping against online harm falls under platforms' responsibility (Schoenebeck and Blackwell, 2021) and is enforced through content moderation, involving both automated and human management of content and user accounts. Content moderation is indispensable for the smooth running of platforms, given the constant and overwhelming flow of content uploaded onto their spaces, which, without it, would be inundated with spam and harmful or violent content (Gillespie and Aufderheide, 2020).

Considering platforms' potential to facilitate or perpetrate harm, content moderation involves a complex interplay of human and technological actors. As a result, platforms are subject to continuous scrutiny while simultaneously managing the flow of information to the public, governments, investors, and advertisers (ibid; Diaz & Hecht-Felella, 2021; Gillespie, 2010). This underscores the critical importance of effective governance and transparent communication in sustaining their operations and profitability (Are & Paasonen, 2021)

Moderation decisions follow in-platform laws known as community guidelines, regulating content related to topics such as violence, terrorism, self-harm, eating disorders, nudity and sexuality (Instagram, n.d.; TikTok, n.d.). However, in their aim to minimise harm, platforms can favour powerful groups and become sources of online and offline harm. For example, internal Facebook documents revealed that the platform's hate speech policy "prioritised protections for white men, neglecting more vulnerable communities" like women drivers or Black children (Diaz and Hecht-Felella, 2021: 8). Following criticism, Facebook promised to deprioritise comments about 'Whites,' 'men,' and 'Americans,' but failed in altering its policies or transparently detailed implementations.

Similarly, TikTok's failure to acknowledge power dynamics when managing hate speech has led to inconsistent and ineffective moderation. This is particularly evident in non-English content moderation, where the platform failed to effectively identify harmful content such as ISIS propaganda in Spanish, slurs against Bosnian Muslims in Serbian (O'Connor, 2021), and the amplification of Russian death threats towards Ukrainian users via viral audio memes (Divon & Eriksson Krutrök, 2023). This highlights TikTok's inability to manage linguistic and cultural nuances, failing to represent the challenges faced by marginalised, oppressed, or at-risk groups.

The absence of a contextual understanding of harm in platform governance often stems from a default, one-size-fits-all approach used globally (Caplan, 2018). For example, while nudity and sexuality encompass complex, necessary aspects of the human experience (Tiidenberg and van der Nagel, 2020), they are often lumped with violence in moderation policies (Instagram, n.d.; TikTok, n.d.), resulting in a disproportionate targeting of sexual and gender expression (Are & Briggs, 2023). This leads to a hostile environment where automated moderation reactions can disproportionately impact marginalised groups, such as sex workers, who already face excessive online abuse and censorship (Gorwa et al., 2020).

Moderation disparities reflect not just platforms' whims but also responses to flawed legislation. For example, the 2018 U.S. FOSTA/SESTA amendments to Section 230 of the Telecommunications Act made social media firms liable for sex trafficking content (Blunt and Stardust, 2021). Their broad definitions, merging sex trafficking (a crime) with sex work (a job), led platforms to excessively censor various users - sex workers, athletes, brands, and educators worldwide - to dodge trafficking facilitation accusations (Tiidenberg and van der Nagel, 2020). As a result, Meta and TikTok often remove accounts that do not actually breach their guidelines, resulting in uneven censorship that over-sexualizes women and LGBTQIA+ users (Haimson et al., 2021).

Platforms' Opaque Practices and Harms

Despite mounting evidence, platforms have denied explicitly targeting marginalised groups with their governance. Thus, to fully comprehend their gaslighting patterns, it is crucial to examine platforms' opaque practices and recontextualize moderation not only as the gatekeeper of external harm, but also as an internal architecture that can itself generate harm.

Unequal moderation decisions often arise from the combination of rules and algorithms created and managed by a largely cisgender, male, heterosexual, white and able-bodied technical workforce, used to moderate content at scale and complemented by outsourced, over-stretched human moderators picking up algorithms' pieces out of context (Diaz and Hecht-Felella, 2021; Gray and Suri, 2019). This means that while platform algorithms may not be designed to make *intentionally* harmful or unfair decisions, their *assemblage* with different layers of platform governance can produce harmful outcomes. Examining assemblages, or "the logics, processes and outcomes of social media content moderation," (Gerrard and Thornham, 2020: 1280) can show how elements like content, interfaces, platform policies, and machine learning can join forces to inadvertently perpetuate sexism. Assemblage theory is relevant in platform miscommunication, highlighting the "silences" in unclear decisions, a familiar issue for users protesting opacity (ibid).

When moderating, platforms can suppress free expression, silencing voices that strive to articulate their communal online and offline vulnerabilities and challenges. Such is the case with BIPOC and LGBTQIA+ communities, ethnic and religious minorities, who often face moderation that involves "downplaying the threats facing them while employing tools and methods that can limit their ability to organize, comment, and document threats to their safety" (Diaz and Hecht-Felella, 2021: 10). Meta's treatment of transgender users is a case in point: the company's own Oversight Board found that while its moderation is quick to remove content featuring transgender self-expression (Oversight Board, 2023), it often fails to moderate severely transphobic content (Oversight Board, 2024). Similarly, on YouTube, creators navigate copyright enforcement through public "copyright callouts," challenging false claims to protect themselves from platform harms such as content removal and income loss (Hallinan et al., 2024).

Typically, the burden of bringing these issues to light has been shouldered by the affected communities themselves, who receive little to no response from platforms (Are, 2023). This has been the case with algorithmic demotion, or shadowbanning. A subtle form of censorship through which platforms algorithmically reduce the visibility of content and profiles by demoting or excluding them from main feeds like the *Explore* or *For You* pages (Are, 2021; Cotter, 2023), shadowbanning is fundamentally characterised by a lack of transparency, secrecy and by platforms' denials of any intervention (Leerssen, 2023) and has been crucial to highlight platform opacity. Indeed, Blunt et al. (2020: 15) argue that shadowbans differ from a ban because "a ban is communicated to a user whereas a shadowban is typically not disclosed to the user."

Shadowbanning results in a reduction of views, under-performing brand collaborations for creators, and the inability to reach new audiences. It is conceptualised as "visibility moderation" on TikTok by Zeng and Kaye (2022: 81), or "the process through which digital platforms manipulate (i.e., amplify or suppress) the reach of user-generated content through algorithmic or regulatory means."

Creators are not directly notified of shadowbanning, leading the term—originally coined in the early 2000s on the *Something Awful* website, though the action dates back to internet forums of

the 1970s—to evolve into a conservative conspiracy theory (Are, 2021). Indeed, although the far-right website *Breitbart* alleged that Twitter was shadowbanning Republicans to diminish their influence in 2016 (ibid), marginalised users have, in reality, borne the brunt of platforms' visibility moderation. For example, pole dancers discovered their hashtags were shadowbanned only when they searched for them, receiving messages that they had been restricted due to community guidelines violations (Are, 2021). Although Instagram issued a direct apology, CEO Adam Mosseri publicly denied the existence of shadowbanning, calling it 'not a thing,' a response criticised as gaslighting (Cotter, 2023) for deceptively dismissing users' experiences.

Mosseri's denial encapsulates a broader pattern of corporate gaslighting, where the pervasive user experiences of shadowbanning are rendered almost invisible due to platforms' lack of notifications, by their weaponisation of algorithmic knowledge and by their governance power (Cotter, 2023; Gillespie, 2022). In response, 'folk theories' (Eslami et al., 2016) have emerged as powerful tools, enabling users to reclaim agency over their visibility, demand platform accountability, and expose systemic injustice. The very need for users to devise and share theories about how curation algorithms work underscores the opacity of platform governance. Indeed, these "communally and socially informed theories and strategies pertaining to recommender algorithms," often referred to as 'algorithmic gossip' (Bishop, 2019: 2602) or 'folklore' (Savolainen, 2022), have ultimately pushed platforms to acknowledge their role in demoting content they consider borderline (Leybold & Nadegger, 2023).

Shadowbanning raises issues of due process, appeal, and resistance, making it a legal matter under the upcoming European Digital Services Act (DSA), which will require platforms to notify users of decisions via "statements of reasons" to enhance transparency and accountability (ibid). Therefore, although platforms are making voluntary transparency efforts to enhance user trust (e.g., TikTok's Transparency Center, which offers insights into content moderation and algorithmic practices, or Instagram's Account Status to verify and contest content restrictions), such initiatives are currently more akin to public relations tools rather than to genuine attempts to improve governance (Burch, 2020; Gerken, 2022). Communication about these processes remains inadequate, as platforms engage in 'discursive depoliticization,' which "promotes individual empowerment and responsibility to avoid acknowledging their own power" (Scharlach, 2024: 5), shifting the burden of avoiding moderation onto users, and further amplifying uncertainties surrounding platform governance.

Shadowbanning exemplifies how users' content, presence, and safety are shaped by unpredictable moderation practices that control visibility on platforms (Gillespie, 2010). However, it is not the only opaque moderation practice platforms utilise. *De-platforming*, the direct removal or banning of users and/or their content from platforms, is another valid example. De-platforming may be triggered by algorithmic detection of content violating a platform's guidelines, but can result from users maliciously or erroneously flagging content as non-compliant with platform rules (Crawford & Gillespie, 2016). This renders *flagging* another non-transparent practice, as platforms currently do not disclose details about the uptake of specific flags (ibid). Despite the intended purpose of flagging, there is growing evidence that it can be misused against certain demographics to harass targeted users and trigger their de-platforming (Are, 2023). Moreover, platforms' notorious over-moderation of nudity, sexual activity, and of violence in public interest scenarios, creates exploitable loopholes for malicious actors to silence groups like sex workers and political opponents during crises. Affected users feel isolated in recovering their profile, rebuilding their network and workplace while having to choose whether to rely on their connections, scammers, or contacts within platforms, or face the loss of profiles, livelihoods, and networks (Are, 2024).

These covert governance interventions fuel what we frame as feelings of gaslighting, rooted in users' experiences of insecurity and mistrust towards the platform's automated decision-making and communication mechanisms. Despite attempts at transparency, platforms' communication efforts about the above governance practices often appear performative and controlled by concealed forces (Gray and Suri, 2019), as accumulating evidence demonstrates that moderation practices can be manipulated to stifle or harass specific users (e.g., Are, 2021; 2023; 2024; Blunt et al., 2020; Cotter,

2023; Oversight Board, 2023; 2024; Silverman and Fortis, 2023). Against this backdrop, we argue that as a systematic discursive strategy, when platforms fail to acknowledge over-moderation or misuse of their tools in both automated and human communication processes, they trigger harmful feelings of gaslighting that are detrimental to their users.

Gaslighting

The term 'gaslighting' is commonly used to describe a form of intimate partner violence (IPV) in which one partner punishes or coerces another whilst denying any wrongdoing, often ridiculing the victim for their beliefs. This process results in a profound loss of epistemic self-trust, eroding a person's confidence in the validity of their lived experiences (Spear, 2019).

Recently the term has been used more broadly to include various contexts characterised by power imbalances in the addressing of complaints. Gaslighting has been felt in healthcare settings, where patient experiences are often overlooked by doctors in favour of 'professional knowledge' (Grim et al., 2019). In academia, women frequently report that their expertise and judgement is undermined, leading to feelings of inadequacy. Similarly, in patriarchal entrepreneurial communities, women entrepreneurs often find themselves isolated and unsupported (Garcia and Martinez, 2019). Omran and Yousafzai (2023) describe how, in Palestine, women entrepreneurs were excluded from business networks and told they were incapable of running successful businesses, with their complaints dismissed as 'overreactions.' Classic gaslighting tactics used against these women included denial, countering (questioning their memories), withholding (refusing conversation), humiliation, and diversion (changing the discussion focus).

The term is also used to describe the denial of institutional racism in the police, despite evidence that non-white citizens have been disproportionately targeted in police searches (Tobias and Joseph, 2020), and in response to politicians and journalists dismissing testimonies as 'fake news' (Rietdijk, 2021). Within the corporate or organisational context, a related concept of 'epistemic injustice' has emerged to describe the dismissal or denial of testimony from individuals. As mentioned earlier, UK postmasters were mistakenly accused of and prosecuted for financial fraud by the Post Office when the faults actually lay with its accounting software system. The postmasters' complaints about the faulty software were dismissed and attributed to individual mistakes, isolating them to prevent the truth from coming to light (Growns et al., 2024).

Whistleblowers in various organisations often experience similar dismissals and discrediting, alongside false accusations of overreacting and lying in their testimonies (Wozolek, 2018). Complaints frequently trigger organisational\corporate gaslighting, where whistleblowers are superficially praised but ultimately portrayed as incompetent or mentally unstable by senior managers (Ahern, 2018). This is also mirrored in acts of organisational social power (Graves and Spencer, 2022), where companies encourage workers to voice systemic issues as a discursive exercise, providing performative attention and care to these complaints without any genuine intention of rooting out the problems.

In all the contexts described, authorities deliberately undermine complainants' testimonies and systematically erode their trust, causing harm arising not just from individual interactions but from extensive communication campaigns characterised by strategic denial of wrongdoing. For instance, a public 'greenwashing' campaign might seek to discredit the testimonies of climate change scientists or policymakers by labelling them as 'fake news,' thereby undermining their scientific credibility (Lopez, 2023). This tactic can escalate to personal attacks, leading to Stern's (2007) stages of gaslighting: disbelief, defence, and eventual withdrawal and self-doubt. Individual harms can be profound, with self-doubt often escalating to PTSD, depression, and suicidal thoughts (Johnson et al., 2021). This was evident in the UK Post Office scandal, where 67% of the wrongfully accused workers reported PTSD symptoms, and 60% experienced depression (Growns et al, 2024).

Similar gaslighting patterns can be observed in platform governance, where human and automated communications *assemble* with processes, policies, and infrastructure that lead to moderation inequalities (Gerrard and Thornham, 2020). These include large-scale moderation, outsourcing content moderation to remote locations often unaware of the content's context, and

community standards rooted in the predominantly North American, puritan, and anti-Arab mentality typical of Silicon Valley (Roberts, 2019; York, 2021). While *each* of these processes may therefore not be put in place to cause harm, *taken together*, the introduction of a series of ad hoc, human, and technical responses to perceived harms without recognizing how these processes assemble becomes harmful to users. This leads to what Duguay et al. (2018: 239) call 'patchwork platform governance,' or "an uneven, often retroactively developed, approach to governance taken by social media platforms, which relies on a combination of formal policies and the selective usage of technological governance mechanisms." Consequently, platforms fail to effectively address or mitigate the harms they generate or abet. The issues and cases of mismoderation arising from this approach are systematically denied by platforms, dismissed as entirely user-generated behaviour, or obscured through generalised comments and bare-minimum minimal acknowledgments (Cotter, 2023).

For content creators, gaslighting can occur when platforms publicly deny shadowbanning despite mounting evidence about it, leveraging opaque algorithmic processes to sow doubt in users' experiences and beliefs (Cotter, 2023), and exploiting the structural information asymmetry among platforms, users, and civil society (Blunt et al., 2020). Indeed, governance techniques like shadowbanning illustrate the information power imbalances that enable gaslighting. Gillespie (2022: 2) notes that social media companies, concerned about appearing biased or unaccountable, often strategically deny or remain circumspect about shadowbanning, with these visibility reduction techniques not being hidden entirely, but allowed to "linger quietly in the shadow of removal policies." These silences may be beneficial for platforms but come at a cost: affected creators must reconsider their content strategy to regain visibility, community, and earnings. This forces them to question their content quality and the loyalty of their followers, while platforms are governing their success algorithmically (Blunt et al., 2020).

It is this strategic gatekeeping and dripping of information that allows platforms to evade scrutiny and avoid the need to divulge business secrets, as highlighted by Patel's (2021) decision to stop accepting quotes 'on background' from Big Tech firms for the site he oversaw, The Verge. He argued that platforms' insistence to drip-feed information to journalists without naming direct sources, preferring ambiguous references to 'a Meta spokesperson', is strategically and intentionally confusing, forcing journalists to rephrase anonymous quotes to influence media narratives without being explicitly accountable.

Given this context, our goal is to analyse platform communication strategies related to platform governance, treating content removals and algorithmic recommendations as distinct yet assembling parts of platform governance (Gerrard and Thornham, 2020; Gillespie, 2022) and delving into the various patterns that constitute platform gaslighting, encompassing the following questions: (1) How do users from Jewish, Palestinian, and sex-positive creator communities interpret both automated and human communications on behalf of platforms (2) How do platforms engage with users about processes and decisions related to moderation and governance? (3) How do platforms articulate their processes and decisions about moderation and governance in public statements about communal issues involving creators that garner public attention?

Methods

Interviews

To address our questions, we adopted a mixed-methods approach, beginning with semi-structured interviews featuring open-ended questions to elicit detailed responses (Roulston, 2021). Our interviews were ethnographic, focusing on participants' descriptions of spaces, events, and actions, supplemented by ongoing analysis of field notes and participation in research settings (ibid). Given our ongoing communications with platform workers, this approach suited our study, shaping both our interview questions and case study choices, as detailed below.

The first author has examined the experiences of 14 Jewish and 10 Palestinian TikTok content creators. Since the Hamas attack on the "NOVA" festival on October 7, 2023, which led to Israel's latest offensive on Gaza, these two user groups have struggled with Instagram and TikTok's heavy and

inconsistent moderation of war-related content. These challenges are compounded by their existing difficulties with platforms' moderation decisions (Farah, 2023).

Recruitment of Jewish creators occurred via a private Facebook group created in 2020, consisting of about 1,000 members who shared their frustrations with TikTok's moderation. The first author access to their group, providing nuanced understandings of their daily experiences and the moderation challenges they face. He then reached out to actively posting members, recruiting 14 Jewish creators with a combined following of over 350,000 from locations including Germany (2), the US (4), Israel (5), and the UK (2), with their social media accounts encompassing themes such as activism, antisemitism, and religion education.

The first author also recruited Palestinian creators through involvement in a peace advocacy network, where Palestinian activists produce content about the Israeli occupation of Palestine on TikTok and Instagram. He learned from network members about heightened moderation issues on TikTok and Instagram for Palestinian-related content since October 7 and also found out that TikTok's Emirates headquarters invited Palestinian creators to a Zoom meeting with the platform's community creators unit to address these issues. Using a snowball approach, AUTHOR 1 then recruited 10 creators from Abu Dhabi (2), Dubai (4) and Palestine (4) who attended that meeting and were experiencing similar moderation challenges. Their accounts encompass themes such as LGBTQIA+ content, foreign policy, human rights, and peace advocacy.

Separately, as part of a broader project on the relationship between flagging and de-platforming on Instagram and TikTok, the second author recruited 12 participants whose accounts encompassed themes such as sex and reproductive health journalism, transgender and non-binary activism, sex work, erotic art, Ukraine war awareness and mental health advocacy. Recruitment criteria reflected her own experiences with de-platforming following viral posts and negative reactions (Are, 2022; 2023) and required participants to have experienced de-platforming of their account and/or content at least once *and* have received at least one negative comment. The recruitment involved a cross-platform call via her network, combined with outreach to users, resulting in participants from the UK (6), Italy (2), the US (2), Ireland (1), and Australia (1), who surpassed 400,000 social media followers, and received £50 each as compensation post-interview. Further solidifying our expertise, the third author had engaged in a study exploring public perceptions of technologically facilitated IPV, investigating perceptions of gaslighting practices.

All authors received separate ethical approval by their institutions, and their interviewees from the three groups were over 18 years of age. They could either remain anonymous, using a pseudonym and an asterisk to avoid further visibility to platforms' moderation, or could be named and credited for their contributions. They were vetted for their experiences ahead of the interviews to check they met the recruitment criteria, and received consent forms and information sheets ahead of taking part. They were interviewed via Zoom, where they communicated challenges related to exposure and content performance, and issues with communicating these to platforms through both human and automated mechanisms.

Platform Responses

Our data includes six notable public responses platforms gave to media and/or civil society organisations about user concerns. These case studies, focusing on issues on Instagram and TikTok, were curated from media sources like Vice, The New York Times, 404Media and official platform tweets for their relevance to community reactions concerning the platforms' lack of transparency. Their inclusion stems from platforms' tendency to present their terms of use as 'community guidelines' and to portray their governance to users as efforts to 'foster community,' which contrasts with the top-down governance approach revealed by these responses.

Our case studies included:

1. Public criticism received by TikTok in 2023 from Jewish organisations for inadequately addressing antisemitism. These criticisms highlighted the ease with which malicious actors

evaded moderation policies to share antisemitic content, questioning TikTok's commitment to the safety of Jewish users (Maheshwari, 2023).

2. TikTok's public response to Jewish users' accusations of favouritism towards pro-Palestine content by attributing the visibility trend to younger users' sympathies and engagement while maintaining the neutrality of its algorithm (Roscoe, 2023; TikTok Newsroom, 2023).
3. Instagram's public reaction to allegations of shadowbanning and censoring pro-Palestinian content, especially posts about the Gaza crisis, resulting in reduced visibility. Meta attributed this to a global technical bug affecting story visibility, a claim met with scepticism by users who suspected targeted censorship of pro-Palestinian content (Navlakha, 2023).
4. Instagram's reaction to backlash for inaccurately translating Palestinian users' bios - e.g., mistakenly translating the Arabic phrase "Palestinian [Palestinian flag emoji] ٱلْحَمْدُ لِلَّٰهِ" as "Palestinian terrorists are fighting for their freedom," causing widespread outrage and prompting an apology from Instagram (Cole, 2023).
5. @InstagramComms' Twitter thread following a 2020 update to Instagram's Terms of Service and a petition by the second author publicly denying that the update was targeting sex workers in a thread. This led sex workers, academics, and creators to highlight numerous instances of the platform disproportionately targeting sex workers (Blunt and Stardust, 2021).
6. Instagram's public comments to their de-platforming of over 50 accounts from London's kink, sex working and sex positive communities, sparking the #StopDeletingUs campaign and a protest at Meta's London headquarters in June 2023 (Smith, 2023). The second author, having mediated between Meta and these communities since 2020 towards swift appeals, was inundated with help requests, leading to the creation of a spreadsheet tracking account deletions and failed appeals. A Meta spokesperson later attributed the removals to an error (ibid).

Analysis and ethical considerations

We analysed our respective interview data independently and shared only anonymised quotes and relevant contexts to protect participant confidentiality. Both interviews and public statements from platforms were analysed via thematic analysis, focusing on spotlighting key answer excerpts identifying, analysing and reporting patterns as our "creative and interpretive stories about the data" produced at the intersection of our theoretical assumptions, analytic resources and the data themselves (Braun and Clarke, 2019: 594).

As thematic analysis aims to provide detailed insights into the lived experiences of users, this method was particularly appropriate for researching stigmatised communities, who necessitate careful consideration of context and depth to fully and accurately convey their experiences (Braun and Clarke, 2019). In our selection of participants' quotes and public responses, we adopted a purposive sampling technique, allowing us to "deliberately look for information-rich cases that capture analytically important variations in the target phenomenon" (Sandelowski, 1995: 81), and to represent the three groups with a cohesive voice.

Findings

Our thematic analysis is organised according to the perceived gaslighting patterns users highlighted: *automated gaslighting*, when platforms' algorithmic processes and accompanying discursive practices invalidate or deny user experience, and *human gaslighting*, when direct communications between users and platforms, or even public comments released by platform workers, also invalidate or deny user experience.

Automated Gaslighting

Our diverse data sets showcased nuanced patterns within platforms' *automated communications* that participants perceived as gaslighting invalidating the adverse experiences they faced on Instagram and TikTok when posting content related to topics like war, hate speech, sex and nudity. The automated patterns we identified are (1) automated responses claiming abusive content did not violate community guidelines; (2) power disparities in content moderation between abusers and their targets and/or between smaller and popular accounts; (3) shadowbanning and algorithmic de-prioritisation of content without notifications; and (4) issues with appealing de-platformed accounts.

No Violation of Community Guidelines

Some of the starkest examples of participants experiencing platforms' automated decisions and processes as gaslighting manifested when reporting hate speech, an action they claim was often followed by notifications that the content was not found to violate community guidelines. For Jewish creators, this situation arose when they reported antisemitic content on TikTok that included references to Holocaust denial, distortion, or trivialization, aligning with the platform's ongoing failure to identify and moderate antisemitic content (McKinnon, 2023):

> "Significant attention was drawn on the 'for you' page to a video from an extremist that juxtaposed a photo of an Israeli affected by recent terror attacks with an image of an oven, alongside the caption 'bake it.' (...). I immediately reported it, but the algorithm failed to flag it for both antisemitism and what appeared to be a clear call for terror. I shared it in our Facebook group, where we share hateful videos with other Jewish creators to mobilize mass reporting and flagging. Astonishingly, we all received a 'no violation of community guidelines' response.'" (Barak_Jewish creator)

> "I often encounter frustrating situations where I report hateful content, such as images of Hitler added with calls for jew killing (...), only to receive moderation responses claiming, 'no violation of community guidelines.' WHAT? HOW? This is frequent with many of my reported videos, and at times, I even questioned whether I was the one that is reading too deeply into hateful videos, but I am not! This is just absurd." (Malka_Jewish educator)

For creators posting nudity, sex work, and promoting sex positivity, such automated minimization affected not only their interpretation of content but also their lived experiences with content abuse. Such was the case of reporting impersonator accounts on Instagram and Facebook while also facing censorship for their own content:

> "I have so many fake accounts using my images and clearly scamming people for money, like either selling my content that they've ripped off of my OnlyFans, or pretending that they are gonna sell my content (...) the amount of times that I reported them over and over again (...) especially one case on Facebook where it was a person called Stephanie something, using all of my photos, in the comments on her photos, on my photos, people are saying, 'I've sent you the money baby,' (...) I reported that account over and over again and Facebook kept sending it back being like, 'We don't find a problem with this account.'" (Reed_Sex educator and sex worker)

In all these cases, platforms denied factual occurrences and stripped creators of their agency, identity, and safety, diminishing their sense of control and adversely affecting their online presence (Are & Briggs, 2023). The simultaneous censorship of participants' own content during their impersonation created a power imbalance where creators felt powerless, while abusers and the platform held the reins on their lives, work, and expression. In these disparate cases of abuse reports, platforms

consistently failed to recognize or seriously address participants' reports of abusive, hateful or illegal content. Although social media companies might consider these instances compliant with their community guidelines, participants were left questioning the legitimacy of their own perceptions in the face of the automated feedback received.

Echoing patterns of IPV gaslighting (Spear, 2019), participants felt discredited and manipulated by the algorithm, which ignored their reports while strictly monitoring their content. This sense of control led them to doubt the validity of their experiences and feel deprived of their agency to report antisemitism and abuse, as platforms continually (re)defined what is considered harmful. Consequently, they felt unsupported by content moderation systems that either ignored or dismissed the impact of harmful content on their well-being.

Our participants' experiences raise further questions about how well moderation-at-scale and outsourced content moderators can perform in situations that require contextual, historical and cultural knowledge. Moderation issues with Holocaust-related content on TikTok contradict its public partnerships against online antisemitism (UNESCO, 2023) and its zero-tolerance stance (TikTok, 2021), especially amid a public outcry over related controversial platform trends (Divon & Eriksson Krutrök, 2023). These issues underscore the gap between TikTok's policies and their implementation, mirroring Caplan's (2018) concerns about the efficacy of *context* moderation, questioning the ability of outsourced moderators to have sufficient sensitivity to historical topics.

Previous research (e.g., Crawford and Gillespie, 2016) already found that an over-reliance on reports allowed platforms to appear to take audiences' views into account while simultaneously deflecting responsibility for their decisions. Our participants' experiences with reporting hateful content on TikTok exemplify the platforms' *reactive* approach to moderation (acting in response to reports) rather than a *proactive* one (actively seeking out and moderating content), outsourcing responsibility and leading to a problematic user-moderation-content dynamic.

Power Imbalances in Content Moderation

The power disparities in content moderation between abusers and their targets and/or between smaller and more popular accounts were another recurring theme. These were acutely experienced by our participants, particularly when the visibility or success of their content was impaired, diminishing their ability to attract audiences and leading them to question the algorithms' inconsistent and unexplained role in recommending posts:

> *"As a Palestinian activist, my focus is on enriching sociopolitical education (...), not monetary gains, and this affects my presence. Despite others producing similar content with triggering imagery related to war and using the same images as me, they often don't face scrutiny from the platform. It's frustrating to see them escape moderation while I have to actively fight for exposure due to the nature of my content. Why doesn't TikTok's algorithm recognize context? It should be taught to do so!" (Bella*_Palestinian activist)*

> *"My video discussing TikTok's hostile stance on fighting antisemitism saw a big drop in performance. (...) The following day, I posted a video addressing antisemitism but not bashing TikTok, and its performance improved noticeably. When I requested a review, I was informed that it would take up to seven work days for feedback, which can truly feel like a lifetime for creators. The review came back, denying any exposure issues (...) I was left unanswered despite clear evidence to the contrary of their claim. Who was looking into my review? Was it even a person? I don't buy it." (Chen_Content creator)*

> *"It's the lack of consistency that's really frustrating to me, that certain people, particularly if they bring in revenue for those corporations, they're allowed stay on but smaller individuals who might actually rely on either having to grow an audience or financially rely on these platforms, they are just coming up against brick wall." (Bel_Transfeminine pole dancing creator)*

Mirroring Duffy and Meisner's (2022) assertion that *size* does matter online, our findings underscore platforms' inconsistent treatment of content creators. This disparity is particularly noticeable among creators posting similarly controversial content who face different levels of scrutiny and visibility. Participants felt that their efforts to connect with audiences were compromised due to uncommunicated internal policies, or shadowbanning and algorithmic demotion influenced by concealed motives, including reputational interests (e.g., Are, 2021; Blunt et al., 2020; Cotter, 2023). Consistently with Gillespie (2022), this heightens the power imbalance between creators and platforms because while removed content is visible for its absence, shadowbanned content leaves no trace, making governance difficult to prove for users and easy to deny by platforms.

References to *bigger* creators, seen as major revenue drivers for platforms, led smaller creators to feel their own stories and experiences were belittled, undervalued and unheard. This dynamic mirrored authoritative gaslighting, such as when police deny systemic unfairness despite evidence of disproportionately targeting marginalised groups (Tobias and Joseph, 2020). Despite the accumulation of instances, users felt a lack of recognition and proper treatment from platform governance decisions (Cotter, 2023), leaving them powerless in their efforts to enhance visibility. Facing *automated gaslighting*, users were uncertain if their content's lack of visibility was due to quality issues or shadowbanning, making them feel as if they were "posting into a void" (Blunt et al., 2020). Creators were preoccupied with expectations about how the algorithm would handle their content, perceiving manipulation on its reach without notification, an assumption supported by their experiments with content and belief in 'algorithmic gossip' (Bishop, 2019).

Moderation Without Notifications

These experiences of isolation, uncertainty and powerlessness led to our third automated gaslighting pattern - moderation of content without notification. Our interviewees expressed frustration with platforms' inefficient processes around reporting abuse, instances of malicious flagging, or recovering deleted accounts. Members of sex-positive and LGBTQIA+ communities learned through connection that their Instagram accounts were de-platformed following targeted action by malicious flaggers in a Telegram group. Their efforts to find solutions to the abuse received through friends and not via Instagram intensified feelings of frustration, highlighting the platform's apparent disregard for user feedback:

> *"We discovered the whole Telegram group issue via acquaintances because there's no real way to tell Instagram that flagging has been going wrong, or that there was no need to flag that specific piece of content, so we asked our friends to send a message we wrote in order to let them know via the Help Centre, but it went nowhere." (Gin_Sex educator)*

De-platforming Appeals and User Isolation

The absence of mechanisms to report behaviours such as malicious flagging, a phenomenon increasingly supported by evidence (Silverman and Fortis, 2023), combined with an automated appeals process, constituted the fourth automated pattern we identified: users feeling isolated and ignored when appealing a de-platformed account. Persistent platform denials regarding practices such as flagging or shadowbanning, exhortations to use the obviously lacking appeals system in these

instances, along with claims of insufficient information about appeal results despite contrary evidence, fostered doubt, frustration, and a sense of isolation among users. The impersonal nature of these responses, largely driven by algorithms, intensified the sense of being gaslit and contributed to a feeling of automated powerlessness (Are, 2022):

> *"I received the same e-mail over and over, saying that I haven't provided sufficient information to show my proof of identity and be reinstated, even though I've sent the picture of my passport to them over and over again, with all my personal information, and showed, 'Hey this is a picture of me on this account, here's a picture of my identification' and then I get the same e-mail telling me that I'm not providing them information they need. And it's not even a human being, it's just an algorithm." (Elia_Transgender sex educator)*

Human Gaslighting

Our datasets revealed distinct patterns in how platform representatives handled user difficulties and complaints, often leaving them feeling that their experiences on platforms were invalidated. These human gaslighting patterns include: (1) absence, inconsistency, and opacity (2) 'band-aid' approaches; (3) dismissiveness towards user experiences; and (4) guarded and cautious responses.

Absence, Inconsistency and Opacity

Key examples of participants feeling gaslit by platform representatives emerged when their regular appeals were ignored, prompting them to seek external help to resolve their issues. Sex-positive users recounted Instagram's opaque decision-making regarding content and profile removal. A lack of transparency around inconsistent appeals processes was exacerbated through human contact which only led to confusion and a sense of greater precarity:

> *"I don't bother doing conventional appeals anymore cause I never hear back (...) the only way that I've managed to get my account back is through people that work at Meta (...) or through management who have direct contacts to Instagram (...) but even then it's been very confusing; this recent time that I've had my account removed (...) I waited two weeks for my account to actually be restored but in that time there was no information and the day before, I got my Instagram account back. I had an e-mail from Instagram saying we reviewed your account and we will not be restoring your account. And obviously that scared the hell out of me, because I didn't know my account was going to be restored the next day." (Reed_Sex worker and educator)*

These inconsistent direct responses to user issues created uncertainty, adding to the already stressful nature of automated moderation (Glatt, 2023). The convergence of user testimonies regarding the platforms' one-on-one communication styles and the way they phrased responses to user concerns mirrors the algorithmic opacity described above (Cotter, 2023). Participants were vocal about the 'elephant in the room,' reflecting Gerrard and Thornham's (2020) notion of 'silence,' as they observed that platforms would often avoid admitting the widely known issues of unpredictability and inconsistency in dealing with users' content. For example:

> *"I'm fortunate to have someone on TikTok. But this support often appears too sporadic, with the TikTok people showing a reluctance to acknowledge the platform's overall arbitrariness with users' content (...) it will cost TikTok more money just to admit that because they will then have to find solutions. Frankly, my human connections with them are just as unreliable as the platform itself. My contact can respond to me within 48 hours, but sometimes it can take up to two weeks. Meanwhile, the harmful content I am trying to draw TikTok's attention to continues to spread. I often wonder if I'm the only one who sees how damaging it is. Why doesn't TikTok prioritize addressing this type of harm?" (Tamar_TikTok eductor)*

Participants frequently noted that it was difficult if not impossible to access platform representatives. They felt neglected, as complaints went unaddressed and as they waited long periods to receive human attention, further blurring the lines between human support and algorithmic responses. They expressed feelings of isolation, exacerbated by the communication styles of platform workers and their automated actions. These experiences not only alienated them but also led them to doubt their own perceptions - akin to victims of IPV gaslighting - resulting in eroded self-trust and diminished resistance to such behaviours (Spear, 2019).

'Band-aid' approaches

Platforms adopted 'band-aid approaches' to governance, merely reinstating user content post-moderation, and issuing apologies whilst failing to address the root problem. Such surface-level approaches to governance would temporarily placate creators, but ultimately left them frustrated as they continued to grapple with the same issues.

> *"I cannot emphasize enough how unhelpful it is that TikTok only provides superficial assistance (...) every time we encounter the same issue, we receive the same inadequate solution and response: 'We are sorry, here are your videos back' along with our apology. This pattern feels like mere lip service. It doesn't seem like they are genuinely sorry; it's more about ticking a box to appear as though they've supported their creator for the day, regardless of the fact that the same creators will just face the same problem again and again." (Yael_TikTok activist)*

Jewish participants shared that TikTok frequently extends a warm welcome through dedicated Zoom sessions where they are encouraged to discuss content moderation challenges, earning them '*the hate cleansers'* title for their active role in combating antisemitism on the platform. However, Jewish users who leveraged their knowledge to produce counter videos addressing antisemitism would often encounter moderation challenges triggered by the algorithm and ignored by TikTok's human moderation team (see Divon & Ebbrecht-Hartmann, 2022). Rather than acknowledging the underlying automation and process-related problem, platform representatives often minimised their support by offering basic guidance to users on how *they* can do better to avoid triggering the algorithm, resembling a practice akin to gaslighters' attempting to shift responsibility onto victims (Johnson et al., 2021; Spear, 2019). Similarly to institutional gaslighting and whistleblowing cases, here, *the hate cleansers* experienced cognitive dissonance: they are commended for identifying harm, yet simultaneously face denial about the extent of the issues they bring to light (Ahern, 2018). This trend was also observed by activist and sex positive communities on Instagram:

> "*[A head of department at IG] had been able to recover these, and made a lot of promises to me and the other organisers about having meetings with higher ups to discuss their violation system and everything. He eventually stopped responding once the protests were over. They even tried to silence us by providing me and a couple organisers with a specific support system, just for us to prevent future take-downs on our account. And I was like, that's not what we're trying to do here - we're trying to get them to fix the system, not just us and a few people. And they didn't do that in the end, so they really gave us absolutely nothing: it was all just empty promises."* (Nellie*_Meme creator)

Dismissiveness Towards User Experience

Users felt most frustrated when platforms dismissed or devalued their knowledge and experience. For example, TikTok downplayed Jewish creators' experiences of hate, despite them being highly visible in their profiles and videos. They testified to observing users with malicious intent exploiting platform features that are rewarded with automatic exposure, like the duet function for video juxtaposition, mobilising others to engage with their videos for antisemitic purposes or inundating their Jewish-themed content with hate comments.

> "*We can tell when our videos become the target of haters and we brace ourselves for a potential dip in performance. It's an unfortunate reality. We don't need TikTok's validation to understand that our exposure is being impacted by the negative comments and mass reporting from trolls. When I shared these concerns with TikTok, the response I received was dismissive, simply stating that video exposure isn't affected by mass reporting. But how can that be? Hundreds of us are experiencing and noticing the opposite. Who are they trying to fool?"* (Dani_TikTok creator)

Platforms' failure to acknowledge the exploitation or misuse of its automated moderation system (Are, 2023; 2024) means that creators feel devalued and begin to question the platform's commitment to transparency and to a fair environment. To fight this, participants joined external communities to collectively challenge TikTok's narratives by gathering and sharing evidence to present back to the platform. However, TikTok's hesitance to recognize these issues, coupled with its demand for creators to document their performance continuously, meant that creators were left feeling unheard, unsupported, and exploited:

> *"Time and again, my contact at TikTok asks me for proof that certain videos are under attack and experiencing reduced exposure. Each time, I gather evidence, such as screenshots of hate comments, the accumulation of these comments, and the slow engagement metrics, in comparison to my other videos that don't revolve around inflammatory topics like antisemitism and islamophobia. Why do I have to put effort not only in creating content for the platform but also in fighting for it to be seen? Despite my attempts, I often receive responses stating that 'based on their data, they cannot definitively conclude if my content is being targeted.' Just sucks."* (Sally_Content creator)

Cautious Responses

A further layer of gaslighting consisted of guarded responses by platform representatives. An example was shared by Palestinian participants who took part in meetings with TikTok representatives to discuss the reduced visibility of Palestinian-related content, especially post-October 7. A group of

users based in the United Arab Emirates approached TikTok's teams in the Gulf region, raising concerns about potential shadowbanning involving specific hashtags. In response, TikTok organised a meeting that included representatives from its community building and trust and safety teams to address these concerns. The participant who attended reported:

> *"(...) Two Palestinian creators within our small UAE-based group had direct links to TikTok. They organized what they called a 'reflective session' for us to voice concerns about the situation in Gaza and our content's performance. Many of us expressed frustration about reduced exposure to our Palestinian-related content. I have a good understanding of my content's performance and can easily spot when something is skewed. My suggestion was that TikTok should reevaluate its approach to certain hashtags as we feel shadowbanned. They were shocked by my assumptions, and their response was dismissive, claiming they don't engage in shadowbanning, they don't favor any sides in this war, and that hashtag visibility is organic and user-driven. I was trying to understand why they were putting words into my mouth with this 'favoring one side' thing, but unfortunately, they repeated their claim about "sides," which left us confused without a clear explanation or solution."* (Noor_Human Rights creator)

TikTok regularly organizes events led by its community building team, specifically targeting diverse user groups who are affected by both online harms and moderation (e.g., Jewish, Palestinian, LGBTQ+, Black communities). While these meetings aim to create a "*safe home for creators,"* they also shape illusory expectations from TikTok: first, to consistently provide an intimate platform for creators to share and alleviate their burdens; second, to foster an atmosphere where creators feel valued and their insights into moderation issues are acknowledged; and third, to uphold a commitment to their protection, as reflected by our participants.

> *"In an online meeting with other Palestinian creators, we were encouraged to expressed concerns about the platform possibly suppressing our videos. Since the attacks on Gaza began, every video I've uploaded has underperformed, receiving little to no engagement, which is not typical for my content. I began to speak in a call with TikTok, where we were allowed to raise concerns. However, the head of the creator community interrupted me, saying, 'I will let you continue with your statement, but please bear in mind that TikTok does not favor sides in this conflict.' I responded that I wasn't suggesting bias but rather wanted to discuss my personal experiences. I'm aware that TikTok is currently on the grill for accusations of favoritism, but their haste to respond was quite suspicious. It's like a guilty child rushing to defend themselves before their wrongdoing is pointed out." (Abaed_Human Rights creator)*

Based on participants' experiences, it appears that TikTok's guarded responses in these meetings align with a public relations strategy, carefully avoiding any acknowledgment of wrongdoing and claiming to be strictly adhering to their policies. This approach includes denying any favouritism on their part, with reactions that reflect the "discursive performances" (Gillespie, 2010) embedded in their community guidelines. It also reflects previous research on corporate strategies to simultaneously praise and minimize whistleblowers' reports (Ahern, 2018; Graves and Spencer, 2022), suggesting that TikTok's interruption of the creator might be perceived as an attempt to dismiss Palestinian concerns by implying their allegations lack foundation. This redirection from users' valid, non-accusatory concerns to other, unrelated accusations, which TikTok aims to disassociate from, appears to be a subtle act of manipulation that is felt as a form of gaslighting, seeking to distort the user's perception of the situation (Omran and Yousafzai, 2023). This dynamic results in a dual layer of doubt concerning both the platform and their own perceptions, leaving participants feeling confused, unheard, and self-doubting (Garcia and Martinez, 2019), starkly contrasting the intended purpose of the meetings, which was to offer a supportive space.

Gaslighting in Platforms Public Response

Platforms' public responses in the media align with what our interviews identified: a strategic choice of words. This *strategy* involves guarded replies or outright denials, attributing issues to algorithmic *glitches*, actively downplaying and minimising user experience. The aim is to depict moderation mishaps as singular occurrences, rather than broader, systemic problems, using the tactic of wielding professional knowledge and opaque language to belittle or deny user experience.

An example can be seen in the removal of numerous accounts related to nudity, sex work, and sex-positive content by Instagram in June 2023 (Smith, 2023), when spokesperson Mitch Henderson gave a guarded response claiming that a "*number of the accounts brought to our attention were removed in error and have been reinstated*." The vague reference to "error" can be seen as an attempt to downplay users' concerns about targeted censorship. By not specifying the nature of the error, the platform's statement obscures the systematic censorship of sexual content and nudity, challenging users' claims that they had been specifically targeted by the platform while keeping platform governance and algorithm processes secret.

Platforms can also dismiss moderation issues without even attempting to conceal it via guarded responses, using open gaslighting, in line with user perceptions of Instagram CEO Adam Mosseri's comment that "*Shadow banning is not a thing*," (Cotter, 2023). The @InstagramComms Twitter statement on their 2020 Terms of Use is a case in point, sharply contradicting the actual terms that ban offering and selling sexual services, affecting even subtle solicitations (Instagram, n.d.). This discrepancy shows Instagram's pattern of denial, clashing with user evidence.

In another example, TikTok faced public criticism from the Anti-Defamation League (ADL) and Jewish leaders for not effectively combating antisemitism. Concerns were raised about the platform's failure to meet its policy commitments, allowing harmful antisemitic content to spread, particularly through misuse of hashtags like #FromTheRiverToTheSea and #StandWithPalestine, some linked to videos promoting the annihilation of Jews in Israel. Responding to the accusations, TikTok's spokeswoman Jamie Favazza emphasised the company's commitment to "*meet with and listen to creators, human rights experts, and civil society to help guide our ongoing work to keep our global community safe.*" However, her statement did little to address the growing disappointment, distrust, and frustration among Jewish creators, particularly when they were told that "*hashtags are created by users, not the platform; it is not the algorithm, it is the users*." (Maheshwari, 2023)

This response not only absolves the platform of any responsibility for platform-perpetrated harm (Schoenebeck and Blackwell, 2021) but also undermines the concerns of creators and patronisingly assumes they don't realise hashtags are user-generated. This tactic evades addressing the core complaints, which centre not on the hashtags themselves but on the *manipulation* of user-generated content and the platform's approach to proactive moderation. Consequently, TikTok's public relations tactics frequently employ a gaslighting approach of victim-blaming (Johnson et al., 2021), overlooking the *custodian* role of platform moderation teams (Gillespie, 2010), sidestepping the company's responsibility in overseeing online content.

Meta's handling of Palestinian content moderation on Instagram, especially after October 7, also involves a strategy of making cautious statements that attribute issues to isolated errors rather than acknowledging broader, systemic moderation problems. After Palestinian creators reported potential shadowbanning of their posts on Gaza's humanitarian crisis, leading to decreased views and engagement, Meta attributed the reduced visibility of stories sharing Reels and Feed posts to a technical glitch, framing the issue as an isolated incident rather than evidence of systemic bias. The company stated the bug affected accounts *globally*, regardless of content theme, and claimed to have fixed it promptly. This response subtly undermines the specific challenges faced by Palestinian users, employing band-aid solutions without addressing the underlying issues.

Additionally, when Instagram was forced to apologise for auto-translating Palestinian user bios into terror-related language (Cole, 2023), Meta failed to respond to journalists, replicating participants' experience of absent or inconsistent communications and failing to acknowledge the hurt and blanket discrimination experienced by Palestinian users.

Meta and TikTok workers' engagement with and responses to users' concerns through both direct and public channels manifest a type of 'epistemic injustice' (Grim et al., 2019) where the imbalance of power between all-knowing tech giants and users, kept distant from the platforms' internal mechanisms, leads to the undermining or outright ignoring of user experience. This highlights the challenge of validating personal stories in tech-dominated spaces, echoing the gaslighting observed by Grim et al. (2019) in healthcare settings, where "professional knowledge" is often systematically dismissive of patient experiences. Social media platforms, as the authoritative figure, effectively weaponise their technical literacy and privileged knowledge of processes against users, whose experiences are strategically placated, dismissed or denied.

Conclusion

This paper frames social media platforms' interactions with users as a discursive communication strategy that leads to the feeling of gaslighting. This occurs when users believe that their valid concerns and complaints are minimised or dismissed by social media companies, ultimately causing them to doubt the validity of their own experiences. Our use of a term rooted in interpersonal relationships is deliberate: social media platforms mediate crucial aspects of our lives, and being excluded from them, and from the opportunities they provide, constitutes both a professional setback and a deeply personal, often traumatic experience (Are & Briggs, 2023). As such, corporate communications strategies that *intentionally* or *unintentionally* result in injustice have profound personal impacts on the individuals platforms claim to uplift, making it essential to emphasise these dynamics through perhaps provocative, yet immediately relatable, analogies. Moreover, as our findings show, platforms repeatedly market themselves to users as 'community builders,' something rooted in and profiting from interpersonal relationships.

Our study demonstrates that platform gaslighting extends beyond users' perception of platforms as "the utmost epistemic authority on their own algorithms" (Cotter, 2023: 1230), as we reveal its pervasive influence across the entire spectrum of moderation practices, blending both automated content moderation and human communication that surrounds it. We expose more nuances to gaslighting, expanding Blunt et al. (2023) and Cotter's (2023) work on shadowbanning to encompass: (1) automated denials of guideline violations for abusive content, (2) moderation biases favouring larger accounts or abusers, especially in instance where platforms' flagging functionalities are misused (3) unnotified shadowbanning and algorithmic suppression, and (4) difficulties in appealing de-platforming decisions.

These practices collectively distort perceptions of platform governance and significantly erode user trust, making it difficult for users to access reliable information or support, providing vague or misleading accounts and attributing systemic problems to isolated errors or to user behaviour. Contextualising our participants' testimonies within current research on platform governance, we highlight how opaque communication often substitutes genuine transparency and accountability. While platforms must manage vast, diverse content and balance the interests of various public and private actors (Diaz & Hecht-Felella, 2021; Gillespie, 2010), we argue that their chosen communication strategies are ineffective and unjust. Thus, whether the denial or minimization of users' experiences is intentional or not, the ambiguity created by these practices fosters a pervasive sense of gaslighting among audiences.

Based on the varied experiences revealed in our data, we illustrate how patterns of *automated* and *human gaslighting* in platforms' interactions with users, reinforced by corporate secrecy and power imbalances, are a key strategy in their communication. We highlighted specific patterns in platform representatives' responses to user difficulties and complaints, characterised by a combination of (1) absence, inconsistency, and opacity in communication, (2) reliance on superficial 'band-aid' solutions, (3) dismissiveness towards user experience, and (4) guarded replies. As a result, a key finding is that our participants do not merely suspect they are being gaslit; they express a clear awareness of it. This is where the benefit of singling out gaslighting lies: a practice that traditionally thrives on doubt (Spear, 2019), gaslighting may hold a measure of agency for those that are able to

identify it. Indeed, our participants described years of accumulated experience, confidently detailing how platforms' opaque handling of their online presence has caused significant harm to them and their communities. This self-assured recognition elevates their claims from mere feelings of manipulation to a well-grounded, documented understanding, showcasing a deeper insight into how platform governance systematically undermines their digital existence.

Our participants felt acutely gaslit in their interactions with Instagram and TikTok. Platform representatives often respond to user complaints about algorithmic bias or attacks by retreating behind policy statements. Users reported that platforms' reliance on official documents to assert the absence of platform-generated harm (Schoenebeck and Blackwell, 2021) amounted to mere lip service, signalled via 'discursive performances' (Gillespie, 2010). Such gestures of concern fail to address the systemic inequalities embedded in platform design, revealing a disconnect between policy rhetoric and user realities. Even transparency efforts, such as directly engaging with users' concerns through dedicated Zoom sessions, prove ineffective due to denials or minimization of the impacts from the assemblage of processes, policies, and technologies (Gerrard and Thornham, 2020) and the 'patchwork platform governance' of moderation actions (Duguay et al., 2018) persist.

Echoing the experiences of gaslighting victims (Grim et al., 2019; Spear, 2019), platform responses can lead users to feel disbelief, defensiveness, and eventually withdrawal and self-doubt (Stern, 2007). Users are kept isolated, their experiences invalidated, dismissed, or undervalued, their beliefs manipulated and their expertise as content creators discredited. Such strategies serve as a shield to maintain monopolies, impede public scrutiny, safeguard companies' public image, and avoid accountability: by failing to provide insights into mistakes, platforms' misconduct becomes unspoken, their authority impenetrable, and their expertise inaccessible. This way, a handful of companies are able to consolidate power and maintain the status quo, leaving users to experience epistemic injustice (Grim et al., 2019), which they struggle to challenge due to the absence of transparent communication about platform wrongdoing.

As researchers exploring the lived experiences of creators, we understand the significance of providing actionable recommendations to platforms, demonstrating how user-centric insights into moderated realities can enhance the value for end users and guide improvements in governance. We have observed that only when creators' well-being aligned with corporate growth incentives, their needs become central to business strategies. Thus, to mitigate users' perceptions of gaslighting and improve their accountability, we urge platforms to prioritise creators in their budget allocations, directing resources toward improving moderation capabilities. This can be achieved by upgrading moderation tools, broadening human moderation teams for more direct engagement with users, and equipping these teams with training on issues sensitive to historical and cultural contexts, such as antisemitism and anti-Muslim sentiments, and differentiating between nudity and sexual content.

To improve human communication, we suggest that platform communications teams move beyond vague explanations that content has been moderated "in error" and instead detail the specific reasons on a case-by-case basis. This approach should include an acknowledgment that repeated errors affecting certain communities may indicate the need for systemic change. Furthermore, platforms should refrain from dismissing user experiences by denying the occurrence of moderation mistakes, opting to investigate reported issues instead. Aside from improving user experience, these changes can be crucial steps for platforms to pre-empt public relations crises that arise when evidence of previously denied errors emerges.

References

Ahern, K. (2018). Institutional betrayal and gaslighting. *The Journal of perinatal & neonatal nursing*, 32(1), 59-65.

Are C. (2021c). The Shadowban cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*. Advance online publication. https://doi.org/10.1080/14680777.2021.1928259.

Are C. (2022). An autoethnography of automated powerlessness: Lacking platform affordances in Instagram and TikTok account deletions. *Media, Culture & Society*. Advance online publication. https://doi.org/10.1177/01634437221140531.

Are C (2023).The assemblages of flagging and de-platforming against marginalised content creators. *Convergence*: 0(0). https://doi.org/10.1177/13548565231218629.

Are, C. (2024). Flagging as a silencing tool: Exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok. New Media & Society, 0(0). https://doi.org/10.1177/14614448241228544.

Are, C. and Paasonen, S. (2021) Sex in the shadows of celebrity. *Porn Studies Forum* 8(4): 411–419.

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.

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. *New media & society,* 21(11-12), 2589-2606.

Blunt, D., Wolf, A., Coombes, E., & Mullin, S. (2020). Posting into the void: Studying the impact of shadowbanning on sex workers and activists https://hackinghustling.org/posting-into-the-void-content-moderation/

Blunt, D., & Stardust, Z. (2021). Automating Whorephobia: sex, technology and the violence of deplatforming: An interview with Hacking//Hustling. *Porn Studies,* 8(4), 350-366.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, 11(4), 589-597.

Centre for Countering Digital Hate. (2022). *Hidden hate - how Instagram fails to act on 9 in 10 reports of misogyny in DMs*. Centre for Countering Digital Hate Inc. https://counterhate.com/research/hidden-hate/

Caplan, R. (2018). Content or Context Moderation? *New York, NY: Data & Society Research Institute*. https://datasociety.net/library/content-or-context-moderation/

Cole, S. (2023). Instagram's Palestinian Arabic bio translation issue. *404 Media*. https://www.404media.co/instagram-palestinian-arabic-bio-translation/

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.

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.

Diaz, A., & 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?ref=welcometohellworld.com

Divon, T. and Eriksson Krutrök, M. (2023). TikTok (ing) Ukraine: Meme-based Expressions of Cultural Trauma on Social Media. In: M. Mortensen and M. Pantti (eds), *Media and The War in* Ukraine, pp. 119-136. Lausanne: Peter Lang Publishing Group.

Divon, T. and Ebbrecht-Hartmann, T. (2022). Youthful Platform Commemoration: TikTok as a Frontier for Holocaust Education and Memory. *Jewish Film & New Media: An International Journal*, 10(2): 231-249.

Duffy, B. E., & Meisner, C. (2023). Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility. *Media, Culture & Society*, 45(2), 285-304.

Duguay, S., Burgess, J., & Suzor, N. (2020). Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence*, 26(2), 237-252. https://doi.org/10.1177/1354856518781530.

Eslami M., Karahalios K., Sandvig C., Vaccaro K., Rickman A., Hamilton K., Kirlik A. (2016). First I "like" it, then I hide it: Folk theories of social feeds. *CHI '16: Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2371–2382). https://doi.org/10.1145/2858036.2858494

Farah, H. (2023, December 20). TikTok moderators struggling to assess Israel-Gaza content, Guardian told. *The Guardian*. https://www.theguardian.com/technology/2023/dec/20/tiktok-moderators-struggling-to-assess-israel-gaza-content-guardian-told

Garcia, E., & Martinez, R. (2019). Gaslighting and imposter syndrome: A mixed-methods study on female academic leaders. *Gender and Education*, 31(4), 515-532.

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, 22(7), 1266-1286.

Gerken, T. (2022, December 5). How to check if your Instagram posts are being hidden. *BBC News*. https://www.bbc.com/news/technology-63907699

Gillespie, T. (2010). *The politics of 'platforms'*. *New media & society*, 12(3), 347-364.

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3). https://doi.org/10.1177/20563051221117552

Gillespie, T., & Aufderheide, P. (2020). Expanding the Debate About Content Moderation. *Internet Policy Review* 9(4). https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy.

Glatt, Z. (2023). The intimacy triple bind: Structural inequalities and relational labour in the influencer industry. *European Journal of Cultural Studies*, 13675494231194156.

Graves, C. G., & Spencer, L. G. (2022). Rethinking the rhetorical epistemics of gaslighting. *Communication Theory*, 32(1), 48-67.

Gray, M., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt

Grim, K., Tistad, M., Schön, U. K., & Rosenberg, D. (2019). The legitimacy of user knowledge in decision-making processes in mental health care: an analysis of epistemic injustice. J*ournal of Psychosocial Rehabilitation and Mental Health*, 6, 157-173.

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), 2053951719897945.

Growns, B., Kukucka, J., Moorhead, R., & Helm, R. K. (2024). The Post Office Scandal in the United Kingdom: Mental health and social experiences of wrongly convicted and wrongly accused individuals. Legal and Criminological Psychology, 29(1), 17-31.

Haimson, O. L., Dame-Griff, A., Capello, E., & Richter, Z. (2021). Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist media studies*, 21(3), 345-361.

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

Instagram. (n.d.). Community Guidelines. *Help.instagram.com.* https://help.instagram.com/477434105621119?cms_id=477434105621119.

Instagram Comms. (2020, December 16). We're seeing some confusion that our Terms of Use update is targeted at sex workers. [Tweet]. *Twitter.* https://twitter.com/InstagramComms/status/1339343799156293632

Johnson, V. E., Nadal, K. L., Sissoko, D. G., & King, R. (2021). "It's not in your head": Gaslighting, 'splaining, victim blaming, and other harmful reactions to microaggressions. P*erspectives on psychological science*, 16(5), 1024-1036.

Klein, W., Li, S., & Wood, S. (2023). A qualitative analysis of gaslighting in romantic relationships. *Personal Relationships,* 30(4), 1316-1340.

Leerssen, P. (2023). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Computer Law & Security Review*, 48 (105790): 1-13.

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

López, A. (2022). Gaslighting: Fake climate news and big carbon's network of denial. In *The Palgrave Handbook of Media Misinformation* (pp. 159-177). Cham: Springer International Publishing.

Maheshwari, S. (2023, December 1). TikTok's C.E.O. Uses Personal Touch to Address Antisemitism Concerns. *The New York Times*. https://www.nytimes.com/2023/12/01/business/shou-chew-tiktok-antisemitism.html

McKinnon, J. D. (2024, January 17). Haley renews call for TikTok ban, echoing charges of rising antisemitism. *The Wall Street Journal*. https://www.wsj.com/livecoverage/gop-republican-debate-alabama/card/haley-renews-call-for-tiktok-ban-echoing-charges-of-rising-antisemitism-ZsOSHNPiF2A4WWmHx1Hl

Navlakha, M. (2023, October 16). People are accusing Instagram of shadowbanning content about Palestine. *Mashable*. https://mashable.com/article/instagram-shadowbanning-censor-israel-palestine

O'Connor, C. (2021). HateScape: Mapping the Online Hate Ecosystem. *Institute for Strategic Dialogue*. https://www.isdglobal.org/wp-content/uploads/2021/08/HateScape_v5.pdf

Omran, W., & Yousafzai, S. (2023). Navigating the twisted path of gaslighting: A manifestation of epistemic injustice for Palestinian women entrepreneurs. *human relations*, 00187267231203531.

Oversight Board. (2023). Oversight Board overturns Meta's original decisions in the "Gender identity and nudity" cases. *News*. https://www.oversightboard.com/news/1214820616135890-oversight-board-overturns-meta-s-original-decisions-in-the-gender-identity-and-nudity-cases/

Oversight Board (2024). Oversight Board Overturns Meta's Original Decision in Post in Polish Targeting Trans People Case. *News*. https://www.oversightboard.com/news/1376420189678927-oversight-board-overturns-meta-s-original-decision-in-post-in-polish-targeting-trans-people-case/

Patel, N. (2021). Updating The Verge's background policy. *The Verge.* https://www.theverge.com/press-room/22772113/the-verge-on-background-policy-update

Rietdijk, N. (2021). Post-truth politics and collective gaslighting. *Episteme*, 1-17.

Roscoe, J. (2023, November 13). TikTok: It's not the algorithm, teens are just pro-Palestine. Vice. https://www.vice.com/en/article/wxjb8b/tiktok-its-not-the-algorithm-teens-are-just-pro-palestine

Roulston, K. (2021). *Interviewing: A guide to theory and practice*. Sage Publications.

Sandelowsk,i M. (1995). Sample size in qualitative research. *Research in Nursing & Health*, 18(2), 179–183.

Savolainen L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *Media, Culture & Society*, 44, 1091–1109. https://journals.sagepub.com/doi/10.1177/01634437221077174

Scharlach, R. (2024). How to spark joy: Strategies of depoliticization in platform's corporate social initiatives. *Social Media + Society*, 10(2). https://doi.org/10.1177/20563051241277601

Schoenebeck, S., & Blackwell, L. (2020). Reimagining social media governance: Harm, accountability, and repair. *Yale JL & Tech*., 23, 113.

Shelby, R., Moon, A. J., Yilla-Akbari, N., Rismani, S., Rostamzadeh, N., Gallegos, J., Henne, K., Nicholas, P., Smart, A., & Garcia, E. (2023). Sociotechnical harms: scoping a taxonomy for harm reduction. *arXiv preprint arXiv:2210.05791.*

Silverman, C., & Fortis, B. (2023). A scammer who tricks Instagram into banning influencers has never been identified. We may have found him. *ProPublica*. https://www.propublica.org/article/instagram-fraudster-ban-influencer-accounts

Smith, S. (2023, June 29). Instagram keeps banning sex-positive and kink accounts, censorship creators. *Dazed Digital*. https://www.dazeddigital.com/life-culture/article/60228/1/instagram-keeps-banning-sex-positive-and-kink-accounts-censorship-creators

Spear, A. D. (2019). Epistemic dimensions of gaslighting: peer-disagreement, self-trust, and epistemic injustice. *Inquiry*, 68-91.

Stern, R. (2007). The Gaslight Effect: How to Spot and Survive the Hidden Manipulation Others Use to Control Your Life. *Chatsworth, CA: Harmony.*

Tiidenberg, K., & Van Der Nagel, E. (2020). *Sex and social media.* Emerald Publishing Limited.

TikTok. (n.d.). Community Guidelines. *Sensitive and Mature Themes.* https://www.tiktok.com/community-guidelines/en/sensitive-mature-themes/

TikTok. (2021). Our commitment to Holocaust remembrance and fighting antisemitism. https://newsroom.tiktok.com/en-eu/our-commitment-to-holocaust-remembrance-and-fighting-antisemitism

TikTok Newsroom. (2023, November 13). The truth about TikTok hashtags and content during the Israel-Hamas war. *TikTok Newsroom*. https://newsroom.tiktok.com/en-us/the-truth-about-tiktok-hashtags-and-content-during-the-israel-hamas-war

Tobias, H., & Joseph, A. (2020). Sustaining systemic racism through psychological gaslighting: Denials of racial profiling and justifications of carding by police utilizing local news media. *Race and Justice*, 10(4), 424-455.

UNESCO. (2023). TikTok joins forces with UNESCO and the WJC to combat denial and distortion of the Holocaust online. https://www.unesco.org/en/articles/tiktok-joins-forces-unesco-and-wjc-combat-denial-and-distortion-holocaust-online

Wozolek, B. (2018). Gaslighting queerness: Schooling as a place of violent assemblages. *Journal of LGBT Youth*, 15(4), 319-338.

Wood, M. A. (2021). Rethinking how technologies harm. *The British Journal of Criminology*, 61(3), 627-647.

York, J. (2021). *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. London: Verso Books.

Zeng, J. and Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. Policy & Internet, 14:79–95.