Information, Communication & Society

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rics20

‘Dysfunctional’ appeals and failures of algorithmic justice in Instagram and TikTok content moderation

Carolina Are

To cite this article: Carolina Are (30 Aug 2024): ‘Dysfunctional’ appeals and failures of algorithmic justice in Instagram and TikTok content moderation, Information, Communication & Society, DOI: 10.1080/1369118X.2024.2396621

To link to this article: https://doi.org/10.1080/1369118X.2024.2396621

Published online: 30 Aug 2024.

INFORMATION, COMMUNICATION & SOCIETY
https://doi.org/10.1080/1369118X.2024.2396621

Routledge
Taylor & Francis Group

'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation

Carolina Are

Centre for Digital Citizens, Northumbria University, Newcastle upon Tyne, United Kingdom

ABSTRACT

This article explores users' perceptions of justice when using appeals on Instagram and TikTok, focusing on the barriers de-platformed users across fields like activism, sex work, sex education and LGBTQIA + self-expression face when using these platforms' automated appeals to recover their de-platformed content and/or accounts. Examining appeals from a platform governance standpoint and drawing from fairness and due process literature, this study finds concerning loopholes within these platforms' appeals, leaving room for discrimination, fraud and scams and leading to user disempowerment. Through interviews with de-platformed users, this paper reveals significant barriers faced by particularly transgender and sex working users when recovering their de-platformed accounts through in-platform appeals. With metaphors of an 'algorithmic cop, jury and judge,' this paper concludes that the needs of marginalised users have been designed out of content moderation and of platforms' processes, leading them to experience the appeals system as opaque, unfair and unjust.

ARTICLE HISTORY
Received 8 November 2023
Accepted 20 July 2024

KEYWORDS
Appeals; Content moderation; Fairness; Algorithms; Content creators ; De-platforming

Introduction

> I have a friend who is a sex worker, and her account was taken down and she found somebody that could get her account back if she paid them. And it sounds like such a huge red flag, kind of scam! But her account was reinstated, and I think this happened to her twice, and she reached out to the same person who got her account back a second time. (Natalie*, Australian artist using Instagram)

This article explores de-platformed users' perceptions of justice when using appeals on Instagram and TikTok, focusing on the barriers those across fields like activism, sex work, sex education and LGBTQIA + self-expression face when using these platforms' appeals function to recover their de-platformed content and/or accounts to explore the widespread problem of faulty de-platforming appeals (Are, 2024b; Are & Briggs, 2023; Myers West, 2018).

CONTACT Dr Carolina Are carolina.are@northumbria.ac.uk Northumbria University, Ellison Place, Newcastle upon Tyne NE1 8ST, UK

Account and content removal, also known as de-platforming (ibid), is part of content moderation, the crucial aspect of platform governance consisting in the deletion and/or censorship of online content, without which social media would be unusable (Diaz & Hecht-Felella, 2021).

A swift, fair and accountable appeals system should complement de-platforming powers, since appeals are a crucial procedural mechanism to determine fairness and accountability in offline and online governance (Common, 2019). Appealing means having the ability to contest 'a decision on content takedowns, account suspension, or account deletion' as an individual user, an action after which social media platforms send the appeals to specific support teams (Klonick, 2018, p. 1648). Normally, upon content or account deletion, both Instagram and TikTok users should receive an in-platform notification and/or an email to notify them of which community guideline they have broken, allowing them to press an 'appeal' button, sometimes with the opportunity to write a comment to the moderator or platform (Instagram, n.d.; TikTok, n.d.). On Instagram, account deletions are often followed by the request to prove user identity through submitting a picture of an ID (Instagram, n.d.).

The appeals process allows platforms to check over potential errors in their decisions, dignifying users and allowing them to participate in decisions that affect them, encouraging consistency and uniformity in the governance process, and allowing communities to interpret and understand regulations (ibid). However, previous research and media reports alike have shown appeals are not always swift or fair, particularly when dealing with content by marginalised communities – e.g., nudity, sex work and LGBTQIA + expression – which is judged more harshly than abusive content and/or content by mainstream users (e.g., Are, 2022; Bronstein, 2021; Diaz & Hecht-Felella, 2021; Kaye, 2019; Smith, 2023; etc.). Scholarly work on content moderation found it disproportionately targets communities at the margins who already face offline stigma, such as LGBTQIA + and specifically transgender users (Haimson et al., 2021), sex workers (Blunt et al., 2020) and disabled sex workers in particular (Coombes et al., 2022). Studies also discovered algorithmic bias against people of colour (Noble, 2018) and highlighted platforms' overall opacity towards users and regulators alike (Are, 2021; Kaye, 2019). Research also addressed the challenges of purely algorithmic moderation (Gorwa et al., 2020) and appeals' lack of due process (Binns et al., 2018; Common, 2019; Kuczerawy, 2022; Schoenebeck & Blackwell, 2021). Creator economy studies focused on Instagram and TikTok as workplaces, highlighting how their precarity and users' reliance on their opaquely and algorithmically distributed visibility affects creators' lives, livelihoods and wellbeing (Are & Briggs, 2023; Bishop, 2019; Duffy, 2020; Glatt, 2022; etc.).

However, studies have yet to focus on the barriers users at the margins face when appealing decisions over their content and profiles, meaning there is scope to evaluate platforms' appeals systems under de-platformed, marginalised users' perspective. Given that a handful of platforms rule over digital space, that they enforce their rule through community standards acting as de facto in-platform laws (Gillespie, 2010) and that one of them, Meta, has created an independent tribunal to oversee its content moderation decisions (Meta, n.d.), this study approaches appeals as a sort of digital criminal justice system where algorithms are the main governance tool enforcing rules, power and punishment in an automated fashion (Schwarz, 2019), to establish whether the process enables education, rehabilitation and justice.

The pitfalls of automated social media governance

Algorithms, or automated content processing and sorting systems (Bishop, 2019), are increasingly being deployed to manage information and content at scale and make important decisions across different fields, from healthcare to human resources, from insurance to the criminal justice system (Marjanovic et al., 2022). Since they reduce reliance on human workforces and therefore costs, algorithms are 'a highly economic governance tool' (Schwarz, 2019, p. 127) that deployed to enforce both traditional and in-platform laws, identifying violators and thwarting suspected violations (Schwarz, 2019). In the criminal justice system for example, different actors such as police forces, courts and parole officers use algorithms for practices such as facial recognition, for DNA profiling, risk assessments and predictive crime mapping (Veale, 2019).

Yet, completely automated decision-making can be dangerous (Duarte et al., 2017), since users at its receiving end have struggled to find mechanisms to hold automated system accountable, to understand or contest their decisions (Are, 2021; Binns et al., 2018).

Like other industries, social media content moderation has become largely automated to manage content at scale (Binns et al., 2018) and has been plagued by similar fairness concerns. Platforms govern through policies, design choices and business models that 'affect the universality and free flow of information on the Internet' promoting or constraining civil liberties (DeNardis & Hackl, 2015, p. 762), engaging in quasi-legal practices such as internal legislation via Community Guidelines or Standards and administering punishment through content moderation (Schwarz, 2019).

Known as 'the detection of, assessment of, and interventions taken on content or behaviour deemed unacceptable by platforms or other information intermediaries' (Gillespie & Aufderheide, 2020, p. 2), content moderation is key area of platform governance which can trigger wide-ranging socio-political effects on global communication and workers' rights (Bronstein, 2021). For example, it can result in financial instability and concerning wellbeing impacts on the users who find themselves shut out of their workplace and networks following governance mistakes or over-enforcement (Are & Briggs, 2023). Objectionable content has become an image issue for social media, who rely on outsourced, commercial content moderation systems, enforced both algorithmically and by underpaid, overworked contractors who have to make split-second decisions over content they are not often sufficiently knowledgeable about, or that they may find traumatic (Gray & Suri, 2019). This means that different layers of content moderation such as demotion, deletion and appeals on Instagram and TikTok are often automated and only rely on human decisions through functionalities like reporting, or upon moderator review (Gorwa et al., 2020).

Previous research has found both platforms' governance to be opaque, lacking transparency and often outright misleading (Are, 2024b; Are & Briggs, 2023; Cotter, 2021). Indeed, automated moderation systems are largely proficient at removing content at scale, but not at educating users about their mistakes (Gorwa et al., 2020; Myers West, 2018).

Instagram and TikTok are private, self-regulating entities that are economically motivated to govern in ways that may benefit their earnings, evading scrutiny by sheltering behind business secrets (Klonick, 2018; Schwarz, 2019). Although they may claim to be committed to respecting the United Nations' Guiding Principles on Business and Human Rights towards diversity, inclusion and the respect of freedom of expression

rights (e.g., Meta, n.d.), the intrinsic economic motivation driving their governance means they cover a unique role as intermediaries providing access to digital relations, workspaces and public debate (DeNardis & Hackl, 2015). Because of this, automated content moderation, which often triggers de-platforming, should be explainable and connected to an effective, fair appeals system. Instead, Instagram and TikTok's de-platforming of users without recourse a significant threat to democracy and work (DeNardis & Hackl, 2015; Klonick, 2018).

Offline, appeals are crucial to any human rights compliant legal system and provide relevant societal benefits even outside judicial reviews (Common, 2019). An effective, culturally, linguistically and politically informed appeals system is therefore essential towards the respect of human rights and due process in content moderation: appeals should be timely, and those appealing should be properly notified and informed of the reasons behind decisions that affect them (The Santa Clara Principles on Transparency and Accountability in Content Moderation, 2021).

Nonetheless, social media users have reported frustration with the way Big Tech companies run appeals, dubbing content moderation as 'dehumanizing' and without recourse (Myers West, 2018, p. 4380). Platforms' reliance on automated moderation and on commercial contractors means that both the removal and the appeals process can result in mistakes due to 'insufficient context-awareness' (Gorwa et al., 2020, p. 10).

Furthermore, sometimes users do not even have a chance to appeal: for example, the moderation of alleged terrorist and/or violent content has been found to disproportionately target Arab and Muslim people, whose content is affected by over-enforcement mistakes (Diaz & Hecht-Felella, 2021). Such is also the case with shadowbanning, a light censorship technique according to which Instagram and TikTok demote from or avoid recommending 'vaguely inappropriate' content to the 'Explore' and 'For You' pages, reducing the visibility of posts and profiles without notifying the user (Are, 2021).

In its current shape, therefore, social media platforms' appeals system seems to exacerbate content moderation inequalities that already see pervasive censorship of particularly marginalised groups such as sex workers, users with disabilities, people of colour, LGBTQIA + account owners and those posting nudity (Are, 2021; Blunt et al., 2020; Coombes et al., 2022; Haimson et al., 2021). These communities have been hit particularly hard by censorship since 2018, on the back of the United States' Congress approval of FOSTA/SESTA. An exception to Section 230 of the US' Telecommunications Act, the Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) and the Stop Enabling Sex Traffickers Act (SESTA) made social media platforms liable for hosting content promoting sex trafficking (a crime) and commercial sex (a job), causing Big Tech companies to over-censor posts worldwide for fear of being accused of breaching the joint bill (ibid). While FOSTA/SESTA has not been found to help fight trafficking or to support its victims, it instead triggered censorship across the globe, in fields ranging from sex work to sport, fashion, sexual health, sex education, activism and self-expression due to platforms' snowballing algorithmic detection and removal of any content remotely visually adjacent to sex trafficking (Blunt et al., 2020; Haimson et al., 2021).

Mitigating the damage caused by FOSTA/SESTA, platform regulation such as the United Kingdom's Online Safety Act and the European Union's Digital Services Act (DSA) seems to be at least in part tackling the issues of transparency and fairness, requiring platforms to act proportionately (Ofcom, 2022), to notify and be transparent with users

about decisions affecting their content in order to facilitate appeals (Leerssen, 2022) to establish independent 'trusted flaggers' to highlight harms and violations (European Union, n.d.), and add remedies for over-moderation in the form of the DSA Articles 20 and 21 (Kuczerawy, 2022). Nonetheless, given the landscape highlighted above, interrogating the functioning of appeals on platforms like Instagram and TikTok is crucial towards limiting unfair censorship's effects on users' lives, livelihood and speech.

Perceptions of algorithmic justice

People's perception of justice can be applied to algorithmic decision-making. Colquitt et al. (2001) divide these perceptions into four main strands: procedural justice, or the process and logic behind decisions; distributive justice, or the allocation of positive and negative outcomes following a decision, and whether they are distributed fairly; interactional justice, or the extent to which those affected by proceedings are treated with dignity and respect by those making decisions; and informational justice, or the explanation and information provided about decisions. These aspects are interconnected, and their relationship with one another contributes to people's understanding of whether decisions made about their situations have been just (Colquitt et al., 2001). Binns et al. apply Colquitt et al.'s rationale to algorithmic decision-making:

> Requiring organisations to explain the logic behind their algorithmic decision-making systems (informational justice) enables affected individuals to assess whether the logic of the system is just (procedural justice), which in turn might moderate their assessments of fairness of the decision outcomes (distributive justice). (Binns et al., 2018, p. 3)

At present, users perceive automated decisions as impersonal and dehumanising, leading some to anthropomorphise algorithms to decipher them (Binns et al., 2018). Indeed, in the absence of clear guidelines, efficient communication channels, and specific information about governance decisions, users have to take matters in their own hands to survive precarious working scenarios: this is particularly crippling on Instagram and TikTok, where users working as professional content creators have to guess and circumvent opaque algorithmic decision-making through exchanging 'algorithmic gossip' (Bishop, 2019), change the wording of their content through 'algospeak' to evade censorship (Steen et al., 2023), and trade in uncertain digital environments where the visibility of their content tied to their earnings is outside of their control (Glatt, 2022). To explain the issues faced by creators, Duffy personified algorithmic systems in the creative industries into 'the algorithmic boss' to account for algorithmic capriciousness (Duffy, 2020, p. 104).

Perceptions of injustice related to algorithmic decision-making have previously been framed as a matter of social and human rights (Common, 2019; Kaye, 2019 etc.), since automated systems and technologies are marginalising many groups of people due to their in-built, societal biases (Gebru, 2020). This emphasis on human rights is already at the heart of the DSA, which recognises platforms' current failures in due process and requires Big Tech to provide effective appeals through internal complaint handling (Leerssen, 2022). Paraphrasing Kuczerawy, we can view an effective appeals system as an online translation of the internationally accepted right to a fair trial, which should aim for procedural fairness, transparency and complaints resolution through qualified human-

decision makers transparency (ibid) and which appeals in their current shape seem to be lacking (Common, 2019).

Therefore, to explore marginalised users' perception of justice Instagram and TikTok's appeals, this article answered the following questions:

> RQ1: How do de-platformed marginalised users perceive the efficacy of Instagram and TikTok's appeal functionalities towards the recovery of their content and/or accounts?
> RQ2: How do de-platformed marginalised users experience content moderation by Instagram and TikTok after successful appeals?

Methods and analysis

To answer the research questions, I carried out 12 virtual, ethnographic interviews with Instagram and TikTok users via the videoconferencing software Zoom. Based on speakers engaging in asking and answering questions, qualitative research interviews see the researcher ask open-ended questions to elicit rich responses from participants (Roulston, 2022). My semi-structured interviews relied on an interview guide to then ask participants for more details through follow-up questions, consistently with Roulston's recommendations.

My interviews were ethnographic and focused on participants' experiences of appealing their de-platformed Instagram and TikTok accounts. Ethnographic interviews rely on participants' description of spaces, actions or events and on researchers' ongoing analysis of data through field notes, observation, participation in research settings (Roulston, 2022). Participants are asked open-ended questions to make sense of their space, time, events, people and activities, and their responses are then analysed qualitatively by the researcher (Roulston, 2022). As is often the case in ethnographic interviews, I participated in the research setting by observing my own ongoing experiences of recovering de-platformed content or accounts on Instagram and TikTok, which informed this research's design, its inclusion criteria, the interview guide's main questions and my fieldnotes (Are, 2022). Instagram and TikTok's crucial role in the creator economy and towards activism (Are & Briggs, 2023), my direct experience with their appeals systems, as well as the networks of participants I gained through them determined my choice to select these platforms towards this paper.

The sample size was determined by the study aim, sample specificity, theoretical background and quality of dialogue (Malterud et al., 2016). Malterud et al. suggest that six to 10 participants are sufficient to describe specific experiences such as those in this paper, given that study's aim is narrow – observing users de-platformed in similar circumstances – and the choice of participants is specific to its objectives. To allow for different experiences, this study stopped at 12 participants, when interviews stopped generating new data.

This paper's data emerged as a separate strand from a connected study on de-platforming via suspected malicious reporting by other users (Are, 2024a). As such, its aim and sample specificity were indeed narrow, prioritising similar experiences of content or account take-downs after suspected malicious reporting, and struggles with account recovery. Therefore, the inclusion criteria specifically required participants to have experienced content or account deletions on Instagram and TikTok *and* have received negative comments on their posts, as both are considered examples that

malicious reporting is taking place, and result in users facing in-platform barriers with automated appeals (Are, 2022, 2024b).

I recruited participants through a blend of targeted outreach to those in my network who had posted about experiencing struggles with appeals, and of onboarding through a cross-platform announcement shared on my Instagram, Twitter and TikTok profiles, which have a combined following of over 400,000. Participants were vetted for their experiences and were only chosen if they met both inclusion criteria. They were paid £50 for their time and expertise after the interviews.

The data collected was transcribed and subsequently analysed through thematic analysis (TA), presenting it via answer excerpts. Characterised by minimal data organisation, TA is a qualitative method which describes data in rich detail and provides insights into relevant lived experiences, making it particularly useful for conducting research on stigmatised communities (Braun & Clarke, 2006). TA allows researchers to identify, analyse and report themes or patterns within data. The researcher develops themes through their own coding, which often reflects data collection questions (Braun & Clarke, 2006).

Consistently with previous studies on the experiences of censored, marginalised users (Are, 2022, 2024a; Are & Briggs, 2023), participants yearned for someone to listen to them, so much that some refused payment for participation because they were 'happy to help,' while others thanked me for giving them space to rant. As a result, the data collected was lengthy, rich and nuanced, analysed through two rounds of data analysis.

This study presents some limitations. Critics of interviews argue that they leave too much room for researcher bias, both in asking questions and follow-up questions and in the framing of answers (Roulston, 2022). Therefore, while choosing participants through direct outreach and a call for participants shared with my network may produce specific results that can be a consequence of legislation that has brought platforms to over-moderate sex-related content (Bronstein, 2021), generating similar experiences of de-platforming, it can also reflect the user demographics in my network of sex-positive, queer followers. Similarly, those who responded to my call can be a reflection of my positionality as a white, bisexual, cisgender woman with experiences of digital censorship, attracting similar follower experiences. Nonetheless, consistently with previous research on the creator economy and de-platforming (Are, 2022; Cotter, 2021), users' experiences are often the only type of information we have about platform processes, making them experts worth featuring in platform governance studies.

The interviews took place between September and November 2022, shortly after the conclusion of a previous round of research carried out as part of the same project to inform interview questions (Are, 2024b; Are & Briggs, 2023). Participants were all over 18 years old and based in the countries where most of my social media following lies: the United Kingdom (6), Italy (2), the United States (2), Ireland (1) and Australia (1). They were: two transgender men posting about their transition journey as well as education and trans rights activism, one, Rob* based in the UK and one, Elia, based in Italy; Diana*, a UK-based cisgender woman journalist and content creator posting about sex education; Nellie*, a US-based cis woman meme creator; Natalie*, an Australian creator and artist; Malli, a UK-based cis man creating content raising awareness about mental health and who has recently also been using they/them pronouns; Cassandra*, a UK-based non-binary model and performer; Marta, a UK-based Ukrainian cis woman creating pro-Ukraine content following Russia's invasion; Bel, an Irish

transfeminine 'wannabe' creator building an audience through pole dance and expressing their identity; Gin, an Italian LGBTQIA + cis woman raising awareness about LGBTQIA + rights and posting about emotional and sexual health; Ale, an Italian, UK-resident non-binary activist using Instagram on a personal and activist basis surrounding topics of LGBTQIA + freedom, Palestine, fat inclusion etc. and Reed, a sex, nudity and sex work advocate, sex educator and podcast host. Participants were chosen because of the similarities of their appeals experiences: although their areas of focus may appear disparate, the data will show how, particularly in content moderation of contentious issues where different factions are at play (e.g., trans rights, the war in Ukraine), securing successful appeals or consistent moderation is challenging for creators. Still, participants' location adds a limitation to this study, meaning experiences of faulty appeals outside of the jurisdictions mentioned are not represented.

Participants were given the opportunity to be named or to be kept anonymous and provided their choice of definitions for themselves and their work. Those who wished to be kept anonymous are referred to through pseudonyms and the asterisk sign. Most of them had been de-platformed at least once, while some had had their accounts disabled up to 10 times, sometimes in a year alone. Some decisions were never over-turned, while others lasted from days to months on end.

In the first round of TA, I placed the data in broader, preliminary themes surrounding malicious flagging, de-platforming and their overall effects on users' lives and livelihoods. In the second round I narrowed themes down to those relevant to this paper, choosing the following three which are then presented through different sections in the findings:

(1) Appeals
(2) Account recovery scams
(3) Shadowbanning and loss of engagement

Results

#1. 'Dysfunctional' appeals

Participants in this study described appealing a de-platformed account as a blend of uncertainty, stress, attempting to speak with an algorithm, relentlessly spending hours appealing, giving up altogether to start a brand-new account from scratch or attempting to reach platform insiders.

Some examples below show how being cut off from an explanation about their de-platforming, and using a system that does not work, resulted in them distrusting the platforms they used for work and self-expression:

> I persistently used all of the options. So, I was emailing, I was going through the app, every option I was doing it. I was copying and pasting the same stuff, and I was relentless, I was like a dog with a bone. And it just seemed that after a while, it started reaching people […]. So I think it's just the utter persistence. […] I think somebody at the office went, 'Oh my God, this guy is doing my head in, go give him his account back' or it finally reached someone that was objective, and they forwarded it saying, let him have it back. (Malli, talking about TikTok)

> I don't bother doing conventional appeals anymore 'cause I never hear back. I have done in the past and I think the only way that I've managed to get my account back is through people that work at Meta that have contacted me or through management who have direct contacts to Instagram. (Reed, talking about Instagram)

Users told similar stories of being cut off from even accessing the appeal function after several instances of mistaken de-platforming, without being able to reach a human at the company to explain the context of their situation. Participants lamented being stuck in 'algorithmic loops' of trying to explain the situation over and over, without any success – they felt they were appealing into a void.

> There just is not enough actual support and it looks like it's all relying on automated email and algorithmic systems that are very dysfunctional. (Nellie*, Instagram)

> Everything is formatted so robotically, so automatically, they're like, 'We understand that this is important to you,' but then they do nothing. (Malli, TikTok)

Participants highlighted significant flaws within Instagram and TikTok's appeal systems, to the point where some, like non-binary activist Ale, were not even able to access the appeal functionality:

> When I tried to do the official appeal, the page wasn't loading. I tried so many times. I think I tried, like, 20 times. And the page was not loading! This is a problem. They're not doing anything for it. I didn't even arrive to the point where I had to upload my document. (Ale, Instagram)

When participants' accounts were apprehended, they were not even served this notice of violation by a visible entity with whom they could communicate – they just witnessed the result of the violation in the form of de-platforming, without being given the chance to explain themselves and feeling forced to use a series of mechanic, unhelpful solutions in the hope to recover their accounts and content.

Appeals were not only time-consuming, glitching and ineffective for participants – they also seemed to outright exclude specific communities, and particularly sex workers and transgender users. While sex workers felt unwelcome on platforms, and found that appealing was often pointless, those who had transitioned but were still waiting for identification documents officialising their identity were forced to upload documents featuring their deadname, only to find that those in charge of appeals did not make the effort to recognise them. Elia, an Italian transgender man, told me: 'The way Instagram sees trans people, transness, trans things, and the way Instagram treats us is discriminatory,' adding that when his profile was deleted, he was asked for ID verification only to hear: 'We can't give you your profile back because you're not this person. Your documents say a name, but on Instagram you're not the same person.'

Participants therefore defined the platforms' appeal system as 'not enough,' inadequate and even 'dysfunctional' or discriminatory – in short, as unjust. In this sense, the success of appeals mechanisms seems to reflect the demographics and dynamics of platform censorship against marginalised communities (e.g., see Haimson et al., 2021) and of malicious flagging against stigmatised users and topics (Are, 2024a, 2024b). In line with Binns et al. (2018), participants therefore perceive appeals to be lacking in informational, procedural and distributive justice, and view them as shrouded in mystery.

Since those interviewed largely found in-platform, automated appeals to be ineffective, preventing them from fairly defending themselves through a fair digital trial (Kuczerawy, 2022) and from account or content recovery, they therefore had to take matters off-platform. Some used email addresses (e.g., TikTok's creator support address) provided by their communities, while others reached out to contacts within Instagram and TikTok, to people with known contacts within the tech giants or to journalists, like Ukrainian activist Marta:

> I had an interview with [a journalist] during those days and I mentioned to him that my account was blocked. On the second day [since the deletion], the journalist messaged me saying, 'Your account is back.' I was shocked that [a journalist] could [get my account back]. (Marta, TikTok)

When participants did manage to recover their account, it was therefore often thanks to friends in high places. Natalie*, an Australian artist, was helped by a Meta connection who 'progressed' her appeal, recovering her account within four days. Similarly, meme creator Nellie* was helped by Meta workers she had met during activism campaigns, who acted on her single case rather than on more systemic moderation issues. She said:

> They even tried to silence us by providing me and a couple of 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 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*, Instagram)

Here, participants seem to be hinting at the inherent soft power within content moderation, with platform workers stepping in to expedite appeals to prevent PR damage only depending on the visibility and/or celebrity of the de-platformed user without actively fixing the systemic issues that have triggered mistaken de-platforming in the first place (Are, 2022; Smith, 2023). Indeed, chances of recovering their accounts were, for participants, largely connected with their existing presence and network: some stated they weren't 'powerful' or 'big' enough to elicit hundreds of complaints about their deletion, leading to account recovery. 'It shouldn't be like that anyway,' Rob*, a UK transgender activist, said in connection with Instagram. 'It shouldn't be dependent on clout.' Yet, even high-profile users covered in the mainstream media and represented by celebrity management teams found automated appeals ineffective.

> This recent time that I've had my account removed, even my manager was confused at the responses. She contacted their contacts at Instagram, and it was a waiting game, so I think I waited two weeks for my account to actually be restored but in that time there was no information. The day before I got my Instagram account back, I had an e-mail saying, 'We reviewed your account, and we will not be restoring it.' 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, Instagram)

The lack of communications between platforms and users therefore did not stop at de-platforming without a warning: it was replicated in notification of account reinstatement too. Users mentioned being told that their account was back by other users, or by journalists who helped them recover it, instead of receiving emails or notifications from platforms. Throughout the process, transparency or, in Binns et al.'s words (2018), informational justice about proceedings, was thoroughly absent.

Automated systems remained an inscrutable power ruling over users' digital lives (Duffy, 2020).

#2 'They are exploiting people who are desperate': faulty appeals leading to scams

Faulty, clunky, late or dysfunctional appeals also open users and their networks up for potential abuse in the form of impersonation, hacking or scams.

Participants reported repeated examples of impersonation resulting from a combination of de-platforming and lack of official, 'blue tick' verification (before this could be purchased). Sex worker and educator Reed mentioned swathes of social media accounts were created through pictures lifted from her OnlyFans account upon her de-platforming, but that her reports for impersonation were not actioned by platforms. Mental health creator Malli, too, recounted the experience of creating humorous content pretending to be an intuitive tarot reader leading to four impersonator accounts trying to scam users for money in exchange of readings. Both participants juxtaposed their experiences of de-platforming with Instagram and TikTok's failure to take up their reports to de-platform impersonators, arguing they felt abused by users *and* not believed by platforms, who would de-platform *them* instead of helping them fight their abusers.

Further, de-platformed participants and their networks found that de-platforming opened them up to hackers and/or scammers who promised to recover their profiles in exchange for large sums of money with no guarantee they would do so:

> When I had one of my accounts taken down a couple years back, I got an image of a menu of how much it would cost […] to get an account restored, and we're talking crazy money here – like it was $1000 or $2000, maybe even $3000 to get an account restored back. And these people aren't legitimate – they just either work in Facebook or Meta themselves or know somebody that does, and will put an appeal through for you. They can't actually predict if you're going to get that back or have that taken down. But they're just sort of doing it on off chance, they are literally exploiting people who are desperate. (Reed, Instagram)

Participants' experiences are consistent with news reporting by Cox (2019, 2021), who found a black market of hackers charging extortionate fees to either get accounts de-platformed or restored. Users' need to work, network and express themselves through social media, coupled with a dysfunctional, discriminatory and ineffective appeals system, therefore creates a breeding ground for scammers and fraudsters, extorting money out of desperate de-platformed users. However, scammers and fraudsters did not stop at de-platformed accounts: they also cast their net out to net their innocent followers, who responded to impersonators in the hope to help their friends who lost their account, or who were tricked into sending intimate images and personal data to those exploiting the de-platforming of high-profile creators.

In this case, lack of informational justice (Binns et al., 2018) did not only affect participants' perception of justice and hinder their chance at a fair trial (Kuczerawy, 2022): it led to direct harms in the form of paying a premium (and potentially being scammed) for a service that platforms themselves should be providing, in order not to be excluded from the important social and financial benefits connected with their digital selves (Are & Briggs, 2023). This added a further layer of injustice to a process that participants already viewed as unjust, something that, as the next section shows, was replicated even in successfully appealed content and profiles.

#3 'Dead in the water': post-appeal moderation

Participants hoped that, once their deleted accounts or posts were restored, they could continue sharing their content with their audiences and potentially reach new followers or customers. However, the effects of de-platforming on Instagram and TikTok seemed to linger on their accounts, leaving a sort of 'criminal record' that, even though technically over-turned through the appeals process, penalised their engagement.

The users interviewed found that, after de-platforming, Instagram and TikTok had removed a set of tools to promote their content and collaborate with other users from their profiles. They found they could not use live, growth or sharing features on Instagram, and that they were banned from sharing links, posting or commenting on TikTok. The journalist and activists interviewed found that even just having videos deleted and later restored would 'kill their engagement' and leave them 'dead in the water' upon restoration, as if they dropped off the algorithm. The loss of engagement shared by users post account and/or content deletion was a constant amongst participants who used Instagram and TikTok for content creation as part of their personal brand. This loss of engagement was often related to the core topics and areas of interest they posted. Indeed, consistently with Haimson et al. (2021), users found that posting specific content – e.g., sex work, sex education, queer content, posting about being transgender or war-related content – resulted in shadowbanning, or the aforementioned algorithmic demotion, to the point where journalist Diana* was told by former TikTok employees that even the term 'sex education' was a flagged term by the platform – and therefore that related content would not perform well or be promoted by algorithms. This added policing users felt after de-platforming was crippling for those who relied on platforms for their main work, who had to rely on testing and on algorithmic gossip to prove something was affecting the moderation of their content. For instance, Italian activist Elia attempted to prove that a shadowban was in place on his account: through methods like those used by previously shadowbanned creators such as Are (2021), he sent content to his most loyal Instagram followers asking to amplify it, only to hear they had not seen his posts for months.

All users interviewed expressed different degrees of frustration with what they perceived as shadowbanning of specific content and of re-uploaded de-platformed posts, and expressed the need and wish for more transparency from platforms:

> I wish TikTok would set up a meeting, even just with sex ed in particular. Let's just have a meeting and talk about it: if I really can't use 'sex' or whatever phrase, can you help me out here? Like, how can we make this work? What can we do to make sure that this content is viewable? (Diana*, TikTok)

Overwhelmingly, therefore, users wished to comply with community guidelines and with platform governance, but found no help, communications or transparency from platforms to allow them to do so, even after they had recovered their wrongly de-platformed account. The inscrutability of platform governance (e.g., Blunt et al., 2020; Duffy, 2020) extended even to successfully appealed content and profiles, perpetrating a cycle of lacking transparency and absent informational, procedural and distributive justice (Binns et al., 2018) that led to an active, unexplained and irreversible criminal record even on successively appealed violations, contributing to a sense of injustice that shone throughout the interviews.

Discussion

The experiences shared by this study's participants show that marginalised, sex working and LGBTQIA + users, as well as journalists and activists, perceive Instagram and TikTok's appeals to be opaque, unfair and unjust, and to be outright failing to allow them to both understand where they went wrong and to have an opportunity for recourse.

Participants showed how these platforms share similar appeals processes and functionalities, but present differences in moderation and appeal uptake. For example, Meta's request for an ID when appealing deleted Instagram accounts provided additional barriers for the reinstatement of accounts. Additionally, while the two platforms presented similarities in the challenges their appeals caused, they differed in the way their processes triggered de-platforming. Algorithmic detection commonly resulted in content or account removal on both platforms, but harassment also played a significant part in censorship: while on Instagram censorship on the back of malicious flagging was often related to users being targeted through private harassment (e.g., after their accounts were maliciously reported by users on the back of their profiles being shared in Telegram groups, or after interactions with fans turned sour and resulted in them communicating to them they were reporting them), on TikTok it was the exposure to virality that led to harassment, comments related to reporting violating content and subsequent de-platforming, consistently with previous work on malicious flagging (Are, 2024a, 2024b). On both platforms however, particularly when it came to the appealing and restoring of content deemed undesirable post-FOSTA/SESTA, these experiences of moderation resulted in faulty, unjust appeals.

Regular mentions to lacking access to a human within platforms, to the robotic and impersonal governance of content that feel so personal and important to participants' individual journey, paints an isolating, unfair picture of platform governance, consistently with Binns et al.'s (2018) participants feelings that algorithmic decision-making is dehumanising. This isolation is a blend of not being able to speak to anyone about their experiences to overturn them, and a feeling of injustice with regards to community guidelines they see applied differently depending on the user, reiterating a lack of due process (Common, 2019; Myers West, 2018).

Findings from this study therefore show that Instagram and TikTok's automated appeals process seems to effectively hinder, instead of helping, participants' chances of recovering their de-platformed accounts, particularly when specific identity or content variables are at play.

Consistently with previous work on perceptions of justice (Colquitt et al., 2001) and algorithmic justice (Binns et al., 2018), participants' perception of Instagram and TikTok's appeals showcased failures in informational justice in the shape of lack of explanation of decisions, leading to lacking procedural and distributive justice, given that the lack of information prevented users from being able to evaluate whether the decisions made on their content and data was just and fair. These users understood their situation as being part of a space where investing in improving the fairness of automated and/or human governance system did not matter, and found existing mechanisms to run appeals impersonal, ineffective and 'dysfunctional'.

Participants' experiences show Instagram and TikTok's appeals system are so far from being fair and accountable that third-party intermediaries like journalists and platform

workers at best, hackers and scammers at worst have to intervene to expedite a glitching, inadequate and clearly under-funded system. This exacerbated offline inequalities by discriminating against and excluding certain groups, triggering precarity amongst those whose work greatly relies on platforms (e.g., Are & Briggs, 2023).

Given that participants perceived their de-platforming and its connected appeals as impersonal and dehumanising, I therefore present their perceptions of justice on and interaction with Instagram and TikTok's appeals in a personified fashion, through the metaphors of an algorithmic cop, algorithmic judge and jury, as well as of a digital criminal record.

In participants' experiences, an algorithmic cop seizes their content and/or profiles through the blunt force of de-platforming, without serving any information as to how they violated broad, loosely and unequally applied in-platform laws that already target the most marginalised (Bronstein, 2021; Gillespie, 2010). Subsequently, an invisible algorithmic judge and jury of one or more moderators – an exploited workforce lacking context over the cases they preside over (Gray & Suri, 2019; Schwarz, 2019) – consider each user's case on Instagram and TikTok, providing judgement and basing decisions on the same unclear, opaque and broad laws. But users do not see the jury, the defence or the prosecution. When they are lucky, they see the decision. If this decision is positive, they may be able to continue their activity on platforms, while still not knowing if they have an active, damning criminal record which, in the case of many of this study's participants, takes the shape of shadowbanning, loss of engagement and key platform features. If the decision is negative, they have no recourse to protest it unless they have friends in high places, but they can find closure with awareness. If, however, they are unlucky, users hear nothing back and have no mechanism to expedite their appeals.

Often, users cannot therefore learn from their mistakes or even *become aware* of those mistakes. In this sense, Instagram and TikTok's automated appeals appear unjust and unfair to users, a failure of justice proving governance is not the same for everyone: users who pay or who have friends in high places are fast-tracked towards account recovery, while those who lack this privilege are left without their account and, therefore, their work or networking tools.

Even more worryingly, after justice is supposedly served – so after innocent content and accounts that were de-platformed are restored – users find themselves still affected by layers of punishment in the form of shadowbanning and lack of visibility, with platforms continuing to apply unjust policing and over-enforcement to their daily social media lives. In this sense, the already weak justice granted by appeals and the overturning of de-platforming – whether that is through official, automated appeals or through account recovery via contacts – is worsened by this unresolved 'criminal record' hanging over de-platformed users, even when their deletions have been found to be a mistake. Therefore, adding to the failure of the traditional justice concept of 'innocent until proven guilty,' which does not stand on platforms who delete first and ask questions later, replicating offline, carceral forms of governance (Schoenebeck & Blackwell, 2021), it appears that users are never fully cleared of their violations, leaving them in a constant limbo as to whether their profiles will be deleted again. Instagram and TikTok users therefore are not only lacking access to a fair trial through a faulty, opaque and unfair appeals system – they have an unresolved 'criminal record' hanging over them forever.

In this algorithmically unjust system there is no room for rehabilitation or even prison: there is just erasure, with de-platforming becoming a blunt force instrument to

erase not just posts, but whole profiles, their connected memories and identities (Are & Briggs, 2023).

Marginalised participants' perception of Instagram and TikTok's appeals as 'dysfunctional' is therefore apt to their situation: the opacity of content moderation makes justice so invisible it may as well be non-existent, since by being invisible it can hardly be questioned. This feeds into the wider opacity of platform governance, which shuns scrutiny to reinforce power monopolies and offline inequalities, withdrawing the right to a fair trial from participants (Kuczerawy, 2022). This has crippling effects on users' freedoms, ability to work and wellbeing, in line with previous research findings determining that deplatforming triggers financial stress, feelings of low mood and shame, particularly against users at the margins (Are & Briggs, 2023).

Conclusion

This article focused on marginalised users' perceptions of justice when appealing de-platformed content and/or accounts on Instagram and TikTok. Their experiences described appeals as 'dysfunctional' at best, 'discriminatory' at worst.

Appeals are 'an essential aspect of a user-oriented moderation process' (Common, 2019, p. 3). While they should provide assurances to all users that content that has been mistakenly removed will be restored and that the rules they are expected to obey are applied consistently (ibid), this study has shown that the algorithmic, automated decision-making governing the appeals system on Instagram and TikTok is a case study in the failure of due process.

Far from only having to deal with the whims of an algorithmic boss (Duffy, 2020), social media users, and particularly marginalised account owners working with and expressing themselves through Instagram and TikTok from the fields of sex education, sex work, LGBTQIA + content, activism and journalism, face a daily stand-off with a wholly algorithmic criminal justice system including an algorithmic cop, judge and jury, and a hanging, indelible criminal record with little to no humans involved, no attempt towards justice, clarity or rehabilitation, or indeed fairness or nuance.

The correct functioning of appeals systems in algorithmic content moderation demands further scrutiny by research and civil society alike, to avoid falling into the trap of accepting platform governance as it stands and relying on the sparse and opaque information platforms provide (Gillespie, 2023). Indeed, while the issue of moderating at scale is crucial and significant, research still needs to question the functioning of algorithmic due process to prevent governance from taking place solely on platforms' terms, without considering user or democratic needs.

While the pressures of governing large platforms are considerable, and the potential harms of the visibility of certain content need to be prevented, it is striking that platform governance has placed such a strong emphasis on detection and removal, but not on the reversal of mistaken decisions, or on the support of those affected by it. Indeed, automated removals and algorithmic detection would likely be perceived as less unfair and crippling if they were accompanied by an equally swift, fair and clear appeals system. However, currently, this is not the case, and algorithmic governance focusing allegedly on keeping users safe is instead not designed with the user in mind, and actively contravenes platforms' alleged compliance with human rights standards and principles.

Therefore, to hold platforms accountable, future research can explore different social networks and internet spaces' appeals systems to compare them with Instagram and TikTok and identify their strengths and challenges. Studies may also consider exploring different user pools' experiences with appeals, and gather data and information directly from moderator teams to shine further light on internal decision-making.

Acknowledgements

I would like to thank everyone who shared and took part in this survey, particularly those who shared personal, traumatic experiences of de-platforming towards advancing knowledge in this field.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

The data that support the findings of this study are available on request from the corresponding author, further information is available at https://doi.org/10.25398/rd.northumbria.26321983.v1. The data are not publicly available due to the privacy of research participants.

Funding

This work was supported by Research Councils UK > Engineering and Physical Sciences Research Council EP/T022582/1.

Notes on contributor

Carolina Are is a platform governance with a PhD in online abuse and conspiracy theories, currently working as Innovation Fellow at Northumbria University's Centre for Digital Citizens. Following her experiences of online censorship, she has been researching on algorithmic bias against nudity and sexuality on social media, and has published the first study on the shadowbanning of pole dancing in *Feminist Media Studies*. Her work has been published in *Social Media + Society*, *Media, Culture & Society, First Monday* and *Porn Studies*, and it has appeared in *The New York Times, The Atlantic*, *The Guardian, The Conversation*, the BBC, *Wired*, the *MIT Technology Review*. She is also a content creator and blogger, as well as an activist and a pole dance instructor, on social media at @bloggeronpole.

ORCID

Carolina Are http://orcid.org/0000-0003-1110-3155

References

Are, C. (2021). The Shadowban cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, 1–18. 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 and Society*, 822–840. https://doi.org/10.1177/01634437221140531.

Are, C. (2024a). The assemblages of flagging and de-platforming against marginalised content creators. *Convergence*, *30*(2), 922–937. https://doi.org/10.1177/13548565231218629.

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

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

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's reducing a human being to a percentage; perceptions of Justice in Algorithmic Decisions. *ACM Conference on Human Factors in Computing Systems (CHI'18), April 21–26, Montreal, Canada*. https://doi.org/10.1145/3173574.3173951.

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

Blunt, D., Coombes, E., Mullin, S., & Wolf, A. (2020). Posting into the void. *Hacking/Hustling*. https://hackinghustling.org/posting-into-the-void-content-moderation/.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bronstein, C. (2021). Deplatforming sexual speech in the age of FOSTA/ SESTA. *Porn Studies*, *8*(4), 367–380. https://doi.org/10.1080/23268743.2021.1993972

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*(3), 425–445. https://doi.org/10.1037//0021-9010.86.3.425.

Common, M. (2019). *The importance of appeals systems on social media platforms*. LSE Law – Policy Briefing Paper No. 40. https://doi.org/10.2139/ssrn.3462770.

Coombes, E., Wolf, A., Blunt, D., & Sparks, K. (2022). Disabled sex workers' fight for digital rights, platform accessibility, and design justice. *Disability Studies Quarterly*, *42*(2), https://doi.org/10.18061/dsq.v42i2.9097

Cotter, K. (2021). "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, *26*(6), 1226–1243. https://doi.org/10.1080/1369118X.2021.1994624.

Cox, J. (2019). *Hacked Instagram influencers rely on white-hat hackers to get their accounts back*. Vice. Available at: https://www.vice.com/en/article/59vnvk/hacked-instagram-influencers-get-accounts-back-white-hat-hackers.

Cox, J. (2021). *Scammer service will ban anyone from Instagram for $60*. Vice. https://www.vice.com/en/article/k78kmv/instagram-ban-restoreservice-scam.

DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, *39*(9), 761–770. https://doi.org/10.1016/j.telpol.2015.04.003

Diaz, A., & Hecht-Felella, L. (2021). *Double standards in social media content moderation*. Brennan Centre for Justice. https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf.

Duarte, N., Llanso, E., & Loup, A. (2017). *Mixed messages? The limits of automated social media content analysis*. Center for Democracy & Technology. https://perma.cc/NC9B-HYKX.

Duffy, B. E. (2020). Algorithmic precarity in cultural work. *Communication and the Public*, *5*(3-4), 103–107. https://doi.org/10.1177/2057047320959855

European Union. (n.d.). Article 19 – Trusted flaggers. Guide to the digital services act. https://digitalservicesact.cc/dsa/art19.html.

Gebru, T. (2020). Tutorial on fairness accountability transparency and ethics in computer vision (FATE/CV 2020). *Google*. https://sites.google.com/view/fatecv-tutorial/home.

Gillespie, T. (2010). The politics of "platforms". *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2023). The fact of content moderation; or, let's not solve the platforms' problems for them. *Media and Communication*, *11*(2), 1–4. https://doi.org/10.17645/mac.v11i2.6610

Gillespie, T., & Aufderheide, P. (2020). Introduction. In T. Gillespie, P. Aufderheide, E. Carmi, Y. Gerrard, R. Gorwa, A. Matamoros-Fernández, S. T. Roberts, A. Sinnreich, & S. Myers West (Eds.), *Expanding the debate about content moderation: scholarly research agendas for the coming policy debates* (pp. 1–29). Internet Policy Review.

Glatt, Z. (2022). We're all told not to put our eggs in one basket: Uncertainty, precarity and cross-platform labor in the online video influencer industry. *International Journal of Communication*, *16*, 1–19. https://doi.org/10.46300/9107.2022.16.1

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). https://doi.org/10.1177/2053951719897945

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop silicon valley from building a new global underclass*. Harper Business.

Haimson, O., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, *5*(466), 1–35. https://doi.org/10.1145/3479610

Instagram. (n.d.). *Learn more about your account*. Help Centre. https://help.instagram.com/384216631681668.

Kaye, D. (2019). *Speech police – The global struggle to govern the internet*. Columbia Global Reports.

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*(6), 1599–1670.

Kuczerawy, A. (2022). Remedying overremoval: The three-tiered approach of the DSA. *Verfassungsblog*. https://verfassungsblog.de/remedying-overremoval/.

Leerssen, P. (2022). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. Forthcoming.

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research*, *26*(13), 1753–1760. https://doi.org/10.1177/1049732315617444

Marjanovic, O., Cecez-Kecmanovic, D., & Vidgen, R. (2022). Theorising algorithmic justice. *European Journal of Information Systems*, *31*(*3*), 269–287. http://doi.org/10.1080/0960085X.2021.1934130

Meta. (n.d). *Corporate human rights policy*. About.fb.com. https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. doi:10.1177/1461444818773059

Noble, S. U. (2018). *Algorithms of oppression*. NYU Press.

Ofcom. (2022). Online safety bill: Ofcom's roadmap to regulation. *Ofcom.org.uk*. https://www.ofcom.org.uk/__data/assets/pdf_file/0016/240442/online-safety-roadmap.pdf.

Roulston, K. (2022). *Interviewing: A guide to theory and practice*. Sage.

The Santa Clara Principles On Transparency and Accountability in Content Moderation. (2018). Santa Clara Principles 1.0. https://santaclaraprinciples.org/.

Schoenebeck, S., & Blackwell, L. (2021). Reimagining social media governance: Harm, accountability, and repair. *Yale Journal of Law and Technology*, *23*, 113–152.

Schwarz, O. (2019). Facebook rules: Structures of governance in digital capitalism and the control of generalized social capital. *Theory, Culture & Society*, *36*(4), 117–141. https://doi.org/10.1177/0263276419826249

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

Steen, E., Yurechko, K., & Klug, D. (2023). You can (not) say what you want: using Algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media + Society*, *9* (3). https://doi.org/10.1177/20563051231194586

TikTok. (n.d.). Content violations and bans. *Support.tiktok.com*. https://support.tiktok.com/en/safety-hc/account-and-user-safety/content-violations-and-bans.

Veale, M. (2019). *Algorithm use in the criminal justice system report*. The Law Society. https://www.lawsociety.org.uk/topics/research/algorithm-use-in-the-criminal-justice-system-report.