Investigating the assemblages of flagging and de-platforming against content creators at the margins

Abstract

This study examines how de-platforming and flagging assemble to replicate offline inequalities, making content creators at the margins vulnerable to both online abuse and censorship on Instagram and TikTok. Highlighting gaps in online harms literature surrounding the misuse of this functionality, this paper frames misused and malicious flagging as online abuse through interviews with users who believed they were de-platformed this way, showcasing this practice's emotional and financial impact on targets and creating a framework to identify it through users' gossip.

Keywords: online abuse, flagging, de-platforming, platform governance, Instagram, TikTok.

This study examines how de-platforming and flagging, two elements of social media content moderation, assemble to replicate offline inequalities, making content creators at the margins vulnerable to both online abuse and censorship on Instagram and TikTok.

De-platforming is the practice of deleting and removing content and/or profiles from an internet platform, a form of governance without which social networks would be unusable (Diaz & Hecht-Felella, 2021). However, social media companies' over-compliance with new platform governance legislation and their attempt to protect their financial and reputational interests by restricting posts they deem unsavoury

has been found to disproportionately affect content picturing nudity, sexuality and sex work (Author, 2021; Paasonen et al., 2019; etc.). Aside from concerning implications for freedom of expression, this de-platforming of content and profiles users rely on for work, for networking, for self-expression and connection has negatively affected users' earnings and wellbeing (Author & Briggs, 2023).

Flagging is an affordance "for reporting offensive content to a social media platform" for it to be de-platformed (Crawford and Gillespie, 2016, p. 411). Similarly to affordances such as liking or commenting, it is a means social media and internet platforms offer users to take action (Graves, 2007) – in this case, to share feedback about the content Instagram and TikTok show their audiences (Goanta and Ortolani, 2021).

Yet, flagging's impact on the de-platforming of content and/or profiles remains opaque due to lacking platform transparency about why or whether user reports have been taken up, or about why content has been reported (Crawford & Gillespie, 2016). Users are often unaware of what triggered their content or account's deletion and lack access to specific information about whether deletions were caused by one or more posts, or by one or more reports by other users (Schoenebeck and Blackwell, 2021). This is particularly true for Instagram and TikTok, which were chosen as research sites because they are two high-profile, free-to-use social media platforms, allowing professional content creators to maintain networks, earn a living through selling products and/or promoting work, and express themselves (Duffy & Meisner, 2022) that, unlike YouTube, do not notify creators when another user has reported their content (Author, 2022; Google, n.d.).

However, research (Author, 2022; 2023; Fiore-Silfvast, 2012), oversight bodies (Oversight Board, 2023), malicious flaggers claiming ownership of reporting

campaigns (Clark-Flory, 2019) and independent journalistic investigations alike (Silverman & Fortis, 2023) have found that the flagging tool has been successfully gamed towards silencing or harassing users that are already vulnerable to platforms' censorship, such as sex workers, LGBTQIA+ and sex-positive users, activists and journalists. For example, conscious of Meta's tendency to de-platform nudity, hackers exploited their networks to blackmail influencers and sex workers, threatening to unleash mass reports on them unless they paid ransom (Cox, 2021; Silverman & Fortis, 2023). Further, Meta's independent oversight body, the Oversight Board, found that as little as three reports sufficed for moderators to remove posts showing partial nudity by trans and non-binary users from Instagram (Oversight Board, 2023). For the Board, users can game the system to remove content that complies with community guidelines, but that they do not agree with or wish to see, taking advantage of Meta's notorious distaste towards nudity: "multiple reports that generate multiple reviews can increase the likelihood of mistaken removals" (ibid). On TikTok, too, reports seem to have influenced the censorship of specific accounts, particularly within the sex-positive, LGBTQIA+ and activist space (see Author, 2022; Perrett, 2021; Stokel-Walker, 2022),

Similarly to how liking and commenting may determine the popularity of content and profiles (Glatt, 2022), flagging therefore has an element of power on Instagram and TikTok. It allows social networks to evade liability and to appear to be conferring power to audiences, potentially handing the already automated reins of platform governance to those who game the system to harass others (Author, 2022; 2023; Goanta and Ortolani, 2021). The misuse or malicious exploitation of flagging has therefore also been defined as organized flagging (Crawford and Gillespie, 2016), user-generated censorship (Peterson, 2013) and user-generated warfare (UGW)

(Fiore-Silfvast, 2012). This misuse can become a form of online abuse resembling "mass vigilantism" (Schoenebeck and Blackwell, 2021, p. 7), where necessary governance functionalities are co-opted by malicious actors to banish users from online spaces (Author, 2023). Examples of misused flagging include: conservative group Truth4Time's coordinated flagging of pro-LGBT Facebook groups; an incel-driven 'purge' of adult performers from Instagram; flagging to de-platform online Cyber-Jihad YouTube videos; (Clark-Flory, 2019; Crawford & Gillespie, 2016; Fiore-Silfvast, 2012).

Users at the receiving end of misused and malicious flagging have felt targeted not just by platforms, but by their own friends, audiences and communities (Duffy & Meisner, 2022; Myers West, 2018). This has opened them and their followers up to further harm – i.e. scams and fraud: Cox (2019; 2021), for example, found that many de-platformed users paid hackers working within platforms to reinstate their disabled accounts without being sure they would actually do so, and that malicious actors also pay hackers to trigger de-platforming.

Previously identified as a tool to aid those targeted with online abuse, platforms' flagging functionality can therefore facilitate abuse instead when it assembles with wider issues within platform governance. However, there is, at present, a gap in research on how flagging can coalesce with an already opaque and unequal platform governance, leaving room to explore its abusive potential. To do so, I carried out 12 semi-structured interviews with de-platformed users who believed they were de-platformed following malicious flagging to describe this practice to help regulators, platforms and users alike to identify it and counter it. Focusing on users from demographics that have been notoriously targeted by censorship – the aforementioned sex-positive, LGBTQIA+ and activist groups - this paper emphasises

the connection between censorship and marginalisation, which in this case covers mainly instances of digital whorephobia, or "the hatred, disgust and fear of sex workers—that intersects with racism, xenophobia, classism and transphobia" (Simpson, 2022: 20), extending to content that is not related to sex work per se, but that is either visually adjacent to it or perceived as adjacent to it, and moderated and stigmatised as such.

This paper begins by analysing online abuse literature, situating online harms within crucial issues in platform governance, such as the approval of legislation shaping the way platforms moderate content. Here, I borrow from Gerrard and Thornham's (2020) notion of social media's 'sexist assemblages' to argue flagging and de-platforming can *assemble*, resulting in online abuse and censorship. Methods and analyses follow, presenting this paper's findings through a thematic analysis. The paper concludes with a framework to identify misused and malicious flagging, which I define as the misuse of platforms' flagging functionality to report content not because it goes against social media community guidelines, but due to personal disagreement and/or lone or coordinated attempts to harass targeted users to trigger their de-platforming. Here misused and malicious flagging are mentioned together to determine different layers of intent in flagging: in the case of the former, the use of the functionality against content that does not violate community guidelines due to personal disagreement on individual posts; in the case of the former, a targeted campaign by one or multiple users against specific content or profiles. To showcase this phenomenon, this paper harnesses the power of user experience, 'algorithmic gossip' (Bishop, 2019) and 'folk theories' (Eslami et al., 2016), or means users have adopted to understand how curation algorithms and content moderation to plan their behaviour in the face of opaque algorithmic governance and that have, in the past,

brought platforms to acknowledge the practice of shadowbanning, or algorithmic demotion (Author, 2021; Lebold & Nadegger, 2023).

Misused or malicious flagging as online abuse

Online harms are influenced by the social circumstances they are enacted in and by those shaping technologies and governance (Castells, 2001). Schoenebeck and Blackwell (2021) distinguish between “platform-perpetrated harms (i.e., those perpetrated by the design of platforms) and platform-enabled harms (i.e., those facilitated by platforms but perpetrated by users or groups)” (p. 8).

Behaviours identified with online abuse include: sending harassing or threatening messages; posting derogatory comments about someone online, or physically threatening or intimidating them; spreading rumours about someone, stalking someone (Hinduja and Patchin, 2007); “threats of rape and other violence, intentionally offensive messages (racist, misogynistic, etc.), hate speech, and libelous personal insults” (Golbeck et al., 2017, p. 1).

Different studies define these behaviours differently, from ‘cyberbullying’ (Hinduja and Patchin, 2007; Navarro et al., 2015) to ‘trolling’ (Keats Citron, 2014) and ‘flaming’, which has often been compared to direct offline harassment (Jane, 2014; Mantilla, 2015). Often used by anonymous users against women in the public sphere and marginalised communities, flaming consists in “issuing graphic rape and death threats,” through a language made of sexually explicit terms such as “homophobic and misogynist epithets,” “scathing, appearance-related judgments” and “coerced sex acts as all-purpose correctives” (Jane, 2014, p. 558, 560). This speech, which Jane defines as e-bile, combines desire and disgust, in a form of

“lascivious contempt” that simultaneously hypersexualises and then derogates targets for their sexuality or appearance (Jane, 2014, p. 560).

Despite the different but overlapping terminologies, previous work on online abuse presents two aspects which are relevant to this study: firstly, the idea of ‘digital mob’ violence against women and marginalised communities; and secondly, the focus on visible abuse, overlooking the misuse of existing harm reduction platform functionalities amongst the behaviours cited as harmful. Misused or failing harm reduction functionalities have meant that anti-discrimination policies often tend to be harnessed by dominant groups (e.g. white men) against protected categories (e.g. Black women), but that the automated tools and the human moderators enforcing them would safeguard those same protected categories when they were harassed (Diaz & Hecht-Felella, 2021). Yet, apart from work by Crawford and Gillespie, Fiore-Silfvast and Schoenebeck & Blackwell, there is little literature on the misuse and the malicious use of flagging, which is both a platform-enabled harm (because of the misuse of an affordance) and a platform-perpetrated harm (due to platforms’ mistaken uptake of malicious flags). It is therefore relevant to situate these practices within the platform infrastructure that enables them, showcasing how different elements of content moderation can assemble to facilitate censorship and online abuse.

The assemblages of flagging and de-platforming

The same existing offline social inequalities and circumstances which influence online abuse also influence the legal and platform frameworks that govern social networks such as Instagram and TikTok, which crack down harder on nudity, sex

and self-expression than on violence, and on marginalised communities instead of their abusers (Author, 2022; Paasonen et al., 2019). While platforms are failing to tackle different types of abuse towards women in the public eye, particularly via direct messages (Centre for Countering Digital Hate, 2022), they have notoriously been over-censoring self-expression, nudity, sexuality and content by those same users after FOSTA/SESTA (Author, 2021; Cotter, 2021; Paasonen et al., 2019 etc.).

An exception to the US Telecommunications Act 1996's Section 230 approved in 2018, the United States' Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) and the Stop Enabling Sex Traffickers Act (SESTA) made platforms legally liable for hosting content that facilitates sex trafficking *and* sex work. Aside from lumping sex trafficking (a crime) in with sex work (a job), and therefore facilitating the de-platforming of swathes of sex workers who lost their livelihoods, networks and opportunities to work in safer, vetted spaces (Blunt & Wolf, 2020; Coombes et al., 2022), FOSTA/SESTA resulted in platforms over-censoring anything remotely sex-related to avoid being accused of enabling trafficking, applying US legislation to global content by triggering censorship outside the US' legal jurisdiction (Bronstein, 2021; Paasonen et al., 2019 etc). Following the exception, even models, performers, sex educators, activists, athletes, lingerie and sexual health brands have had their content censored and their profiles disabled highlighting the influence of this US government policy on international platforms (Author, 2021; 2022; Cotter, 2021).

Since content moderation takes place in a complex ecosystem of policies, algorithmic recommender and demotion systems, automated detection tools and human enforcement (Bishop, 2019; Diaz & Hecht-Felella, 2021; Goanta and Ortolani, 2021), it becomes relevant to frame flagging and de-platforming as

moderation *assemblages*, or “the logics, processes and outcomes of social media content moderation” (Gerrard & Thornham, 2020: 1279-1280). Different elements of the running of platforms, which Caplan and Gillespie (2020) call 'tiered governance,' can blend with each other creating a patchwork of different results, meaning the notion of assemblages can become particularly useful when discussing how they come together to perpetrate online abuse and reinforce offline inequalities against marginalised groups. Used by Gerrard & Thornham (2020) to conceptualise how different elements in platform governance come together to perpetrate sexism, assemblages are the human and mechanical elements joining forces to generate specific results. As such, interrogating how different elements of content moderation – and in this case flagging and de-platforming – assemble means asking questions about content, interfaces, platform policies and machine-learning to understand what the joining of different elements of platform governance does to the social world (ibid).

In the case of misused and malicious flagging, the content is nude, sexual and queer posts disproportionately targeted by platforms post-FOSTA/SESTA (Author & Paasonen, 2021). The platform policies are rules, which are set subjectively, informed by the rule-setters’ biases, or, as Roberts (2019) argues, made “in the specific and rarefied sociocultural context of educated, economically elite, politically libertarian, and racially monochromatic Silicon Valley, USA” or in socially conservative China (pp. 93–94). As a result, platforms’ machine-learning processes are informed by and targeted towards enacting this worldview, disproportionately regulating content by marginalised communities while falling short of protecting them and letting harms by dominant groups go unpunished (Diaz & Hecht-Felella, 2021). This means that misused and malicious flagging of specific content – and particularly

nude, sexual and LGBTQIA+ content (e.g. Crawford and Gillespie, 2016; Clark-Flory, 2019)– can influence content moderation by way of de-platforming, enacting a sex-averse, puritan platform governance characteristic of a post-FOSTA/SESTA internet (Paasonen et al., 2019).

Using assemblages theory to understand the misuse of flagging is relevant because “assemblage theory helps to direct us to silences” (Gerrard & Thornham, 2020: 1280), and flagging is characterised by silence: platforms share very little information about its use or uptake (Crawford & Gillespie, 2016), and users only see the action’s result, such as de-platformed content or profiles, often after different forms of online abuse or of virality (Author, 2022; 2023; Perrett, 2021; Stokel-Walker, 2022).

The misuse of flagging on Instagram and TikTok and its relationship with de-platforming demands more scrutiny to interrogate the effectiveness of platform affordances, highlight user experience and educate social media account owners about a still under-researched form of online abuse. Towards these aims, this paper answers the following questions:

RQ1: How can misused or malicious flagging towards users’ content and/or accounts be identified?

RQ2: How does de-platforming after suspected misused or malicious flagging affect targeted users?

Methods

For this study, I carried out 12 semi-structured, virtual, ethnographic interviews via the videoconferencing software Zoom. In qualitative research interviews, the researcher asks open-ended questions to elicit rich responses from participants (Roulston, 2022). Ethnographic interviews rely on participants' description of spaces, actions or events and on researchers' ongoing analysis of data through field notes, observation and participation in research settings (ibid). The researcher begins by asking participants open-ended questions to make sense of their space, time, events, people and activities, to then qualitatively analyse the data (ibid). Researchers carrying out ethnographic interviews often participate in the settings they study (ibid), which I did by observing and noting down my own ongoing experiences of censorship following perceived flagging on Instagram and TikTok – experiences often similar to participants' and which informed this research's design, its inclusion criteria and interview questions (Author, 2022).

The study's aim, sample specificity, theoretical background and data determined the sample, since particularly when the study's aim is narrow and the combination of participants is specific to its objectives, Malterud et al. (2016) argue it does not require large participant numbers, as the data produced will be relevant and rich.

This study's aim and sample specificity were indeed narrow, prioritising similar experiences, i.e. content or account take-downs after perceived malicious flagging. I therefore capped participants at 12 to diversify the represented experiences to include both users posting content that has been notoriously over-censored (e.g. nudity, sex work, LGBTQIA+ expression, as shown by Author, 2021 and Haimson et al., 2021) and journalists and activists (Stokel-Walker, 2022).

To take part, participants had to have experienced content or account deletions on Instagram and TikTok *and* have received negative comments on their posts, mirroring my experiences of account deletion after receiving swathes of negative comments upon going viral (Author, 2022) and other instances where flags triggered de-platforming (Clark-Flory, 2019; Perrett, 2021; Silverman & Fortis, 2023). Recruitment took place via a blend of targeted outreach to users in my network who shared similar experiences, and of onboarding through a cross-platform call for participants shared across my Instagram, Twitter and TikTok profiles with my combined following of over 400,000. Before confirming their participation, participants were vetted for their experiences and were therefore chosen only if they met both inclusion criteria. Following the interviews, they were paid £50 for their time and expertise.

The data collected was transcribed and subsequently analysed through thematic analysis (TA) in the form of answer excerpts. TA is a qualitative method which allows researchers to identify, analyse and report themes or patterns within data (Braun & Clarke, 2006). Themes are “creative and interpretive stories about the data, produced at the intersection of the researcher’s theoretical assumptions, their analytic resources and skill, and the data themselves” (Braun & Clarke, 2019, p. 594). The researcher develops them through coding, often reflecting data collection questions (ibid). Characterised by minimal data organisation, TA is a realist method describing data in rich detail and providing insights into relevant lived experiences, particularly when conducting research on stigmatised communities (Braun & Clarke, 2006).

Consistently with my previous experiences of carrying out studies amongst censored, marginalised users (Author, 2023; Author & 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 told me: “Thank you for giving me the space to rant.” The data collected was therefore rich, lengthy and nuanced, leading me to conduct a preliminary analysis, to then analyse the responses collected through two rounds of data analysis.

This study presents a set of 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). And indeed the choice of interviewees, both amongst those to whom I directly outreached to and those who responded to my call for participants, can be both a consequence of legislation that has brought platforms to over-moderate anything mildly sex-related (Paasonen et al., 2019), generating similar experiences of de-platforming, but also a reflection of the user demographics in my network of sex-positive, queer followers, with whom I engage from the positionality of a white, bisexual, cisgender woman with experiences of digital censorship.

Critics of my approach may argue that trying to provide evidence of malicious or misused flagging leading to de-platforming through the experiences of de-platformed users may not be effective, as it may be relying on conjectures or misunderstandings of content moderation processes. Some may wish I would have spoken to those who engage in flagging instead of to those affected by it, or argue that those reporting participants may not be doing so maliciously, and may just be expressing their disagreement with content. However, as the data will show, participants’ content *often complies with public-facing community guidelines*, showing discrepancies in platforms’ enforcement. Further, in the face of platforms’ known lack of transparency (Author, 2021), flagging either manifests as a lack of

change on participants' profiles or on their de-platforming: it is therefore an action difficult to detect from outside platforms, a silence (Gerrard & Thornham, 2020), making its manifestation from a user's perspective relevant to consider. Lastly, and most importantly, following a series of investigations (Clark-Flory, 2019; Cox, 2021; Silverman & Fortis, 2023) and Meta's own Oversight Board's (2023) findings that repeated reports resulted in the deletion of non-violating content, and since investigating similar conjectures has shed light on equally mysterious moderation practices such as shadowbanning (Leybold & Nadegger, 2023), listening to participants' experiences, conjectures and suspicions is crucial in platform governance. Indeed, consistently with previous research on creator communities and de-platforming (Author, 2022; Cotter, 2021; Garcia-Rapp, 2017), the specific experiences of creators are often the only type of information we have about platform processes, making users experts in their field. Since Big Tech tends to discredit and patronise participants, I wish to instead believe them, and consider them as collaborators in holding platforms to account instead of entertaining platforms' black box gaslighting about their governance (Cotter, 2021).

Analysis

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). Interviewees were given the choice to be kept anonymous (and therefore protected from further harms and referred to with a pseudonym and an asterisk) or to be named and credited for their experiences and expertise.

Each participant provided a preferred description of their accounts' aim and niche, which this paper has attempted to follow. They were: two transgender men working as writers posting about their transition journey as well as education and activism - one, Rob*, based in the UK and one, Elia, based in Italy; Sylvia*, a UK-based cisgender woman journalist and content creator posting about sex education; Nicole*, a US-based cis woman meme creator; Roxie*, an Australian creator and artist; Malli, a UK-based cis man creating content about mental health; Valery*, a UK-based non-binary model and performer; Marta, a UK-based Ukrainian cis woman creating content about the war in Ukraine; Bel, an Irish transfeminine 'wannabe' creator building an audience through pole dance and expressing their identity; Gin, an Italian LGBTQIA+ cis woman posting about emotional and sexual health; Ale, an Italian, UK-resident non-binary activist using Instagram on a personal and activist basis and Reed, a sex, nudity and sex work advocate, sex educator and podcast host. Most participants had been de-platformed at least once, with some claiming to have had their accounts disabled up to 10 times, sometimes in one year alone. Some deletions were never overturned, while others lasted from days to months on end. Only one user's account had never been de-platformed, but had different types of content removed.

The first round of TA consisted in placing the data in broader, preliminary themes surrounding misused flagging, de-platforming and their overall effects on users' lives and livelihoods. The second round narrowed themes down to three:

1. Malicious or misused flagging in action
2. Cross-platform abuse and lack of nuance in moderation
3. Malicious flagging and de-platforming's offline impact on users and their circles

The following sections present the themes through participant quotes and the researcher's sense-making, with the aim to show how de-platforming after malicious flagging manifests.

Malicious or misused flagging in action

All participants in this study believed they had been targeted with malicious flagging, and provided significant examples of the practice in action. However, their relationship with this functionality was conflicted: while they described flagging as an action platforms allowed users *to perform*, participants also viewed it as something *done to* their account rather than something they themselves could do, because often when they did it, platforms would not listen.

Participants identified a blend of negative comments after their posts had gone viral, the disappearance of older posts and the accumulation of community guidelines violations for disparate reasons unrelated to their behaviour as examples that flagging was being weaponised against their account. For instance, transgender activist and educator Elia became suspicious he was being reported by other accounts when "a lot of posts and stories that were getting deleted didn't have anything explicitly against community guidelines" (i.e. they were not nude or explicit), resulting in a barrage of deletions of even older posts, when his profile was personal and not professional. Users connected misused or malicious flagging initially to a series of 'punishments' (i.e. removal of the ability to go live, to comment on or save posts, followed by account deletion warnings) and then to full de-platforming.

The experiences shared by participants highlight major flaws within platforms' flagging affordance, ranging from platforms triggering de-platforming after flags by users reporting others as a mode of feed curation, to remove content they did not wish to see, all the way to maliciously reporting others according to completely unrelated community guidelines violations to silence creators. E.g. see the experiences of Elia, of Ukrainian activist Marta, of transfeminine 'wannabe' creator Bel, and of mental health advocate Malli:

- *"I got the notification from Instagram that my post was against community guidelines and that it had been removed, sometimes for things that really astounded me, such as sex acts or violence against minors."*

- *"I had a video that got 7 million views taken down that was not violent - it was literally filming of interviews with Putin - but I think because a lot of people in the comment section were Russians, I think a lot of people reported it. So it got taken down permanently and they said it's impossible to reinstate it."*

- *"You can tell that it's just bigots that are just blind reporting something that they don't like to look at, because I've had videos reported under every category. Every type of content moderation flag that TikTok has, a video of mine in some way or form has been put up against it."*

- *"I was doing a live and somebody reported that for sexual activity and nudity, and I got shut down for four days. It means no one is looking at it,*

when something is reported for nudity and sexual activity when there isn't either in the video. [..] It's completely algorithmic, which is an easy way to do things - it's not efficient or fair."

As shown by the above quotes, participants use algorithmic gossip (Bishop, 2019) or folk theories (Eslami et al., 2016) to make sense of their experiences. This attempt at sense-making happened throughout data collection, and can be interpreted as an example of the confusing nature of opaque content moderation rather than as an attempt by participants to justify their actions. And indeed, this discrepancy between the content users posted and the community guidelines according to which it or their profile was removed seems to lend legitimacy to their own folk theories (ibid) or sense-making: it seems odd that mistaken detection could be the only reason behind this censorship, particularly in instances when they were not posting any new content. Instead, it seems that platforms' flag function may be *assembling* with their decision to de-platform content and profiles they already deem controversial based on the assemblage of existing prejudice against content, platform policies, machine learning and enforcement (Gerrard and Thornham, 2020), in a similar vein to findings by Silverman & Fortis' (2023).

At present, the number of flags required for content or account removal is not publicised by platforms. Users receive confirmation that their account is being reported only when platform workers or users tell them about it. After her personal account was de-platformed, sex worker and educator Reed received a message from a Facebook worker on her podcast's official account:

“This girl, when she looked into it she said it looked like there were a couple of accounts targeting my account and constantly reporting posts on my account, up to a point where it basically filled up the quota that Instagram needed for it to be automatically taken down.”

Flagging creates power imbalances between followers and users, who repeatedly found that content requiring the application of nuance and context to understand the meaning of posts often became a way in for malicious flaggers. For example, participants shared that some of the content they believed triggered flagging and de-platforming included: the words ‘bitch’ or ‘fag’ used by women and queer participants removed as hate speech; humour by white people towards white people removed as hate speech or bullying; war reporting removed as violence and graphic content; sex education mistakenly removed as nudity and sexual activity. Users have therefore had to negotiate what they could and could not post, or even which type of content they chose to try to recover: no swearing, no close-ups of anything beyond their face, not even Renaissance paintings showing nudity.

A blend of algorithmically, automatically censored content and posts flagged by others, taking advantage of algorithmic and moderation loopholes within Instagram and TikTok, have therefore resulted in post or profile take-downs for participants. Users can effectively game the system, taking advantage of the assembling of unclear, puritanical and broad community guidelines, automated moderation and underfunded, under-paid human moderators to attempt to remove content and profiles they disagree with. This is particularly evident with nuanced but potentially controversial content, such as sex education, LGBTQIA+ content, nudity and war reporting, in an algorithmic rendition of Jane’s “lascivious contempt,” simultaneously

abusing and frowning upon women and marginalised communities (Jane, 2014: 560). Yet, despite known controversies about this content (e.g. see Perrett, 2021; Stokel-Walker, 2022), platforms do not seem to have added any layer of protection to detect nuance in flags or understand their context. Indeed, it appears that a report by an anonymous account set up for harassment or lurking has the same relevance as any user looking to report concerning content.

Participants also shared that platforms lack specific functionalities to signal that malicious flagging may be taking place. At present, there is no 'misuse of flagging' specific 'report abuse' category beyond 'bullying and harassment' to highlight that rogue activity is taking place, and in a scenario where hate speech and harassing comments are not demoted or removed by platforms, users find it even more challenging to report accounts for misuse of flagging as bullying. Malli highlighted that, due to platforms' lack of vigilance, flags could come from the same person creating multiple anonymous accounts: "People are afforded a certain amount of anonymity on it, because there's so many accounts that don't have anything, no picture, no information, and each one of those is potentially a troll." Additionally, trans activist Rob* shared that the only way he could secure a take-down for harassment and coordinated flagging by a known anti-trans activist was by reporting the pictures she lifted and reposted from his account as a copyright violation.

In this sense, users feel targeted from multiple fronts: from users who mistakenly report content they disagree with; from harassers who engage in visible abusive behaviours through comments and, supposedly, coordinated flagging efforts; and from platforms, who consider these efforts as valid without considering the context, and without providing users with an explanation or a form of complaint for what is going on, showcasing a combination of platform-enabled and platform-

perpetrated harms against the most marginalised users (Schoenebeck & Blackwell, 2021), and the assemblage of different layers of platform governance (Gerrard & Thornham, 2020), perpetrating censorship and online abuse. Rob*'s response signals a form of protest against these multiple attacks: it shows users resort to utilising alternative governance mechanisms as a form of resistance to make up for lacking platform protection and to fight the assemblages of flagging and de-platforming.

Cross-platform abuse and lack of nuance in moderation

Online abuse via misused flagging escalated when it became a coordinated cross-platform action. Indeed, although this study focuses on Instagram and TikTok, a variety of participants experienced cross-platform abuse: they were targeted by users on one platform and then the abuse spread to other platforms and offline. This included harassment and accusations (e.g. transgender users being accused of paedophilia) and flagging leading to content and account take-downs on their main content creation platforms Instagram and TikTok following specific interactions with other users or after viral posts.

Outside of Instagram and TikTok, users found that malicious actors mainly used Twitter and Telegram to coordinate cross-platform abuse to harass and de-platform them. Although based in different countries, this study's transgender participants, for instance, shared similar experiences: Rob* found that anti-trans Twitter users from the UK would periodically go through "all prominent trans people's accounts [on social media] and find things to report," While Elia, based in Italy, said:

"Through a friend who sent me a screenshot I noticed the existence of a Telegram group – which was triggering in itself because it was called something like, 'Let's off them all.' […] The screenshots I was sent were referring to my posts, so people were sharing screenshots of my posts in the Telegram group saying, 'Go flag this person,' as well as sharing a series of slurs. All the screenshots were of my face, or just of me, so they weren't criticising my work, or the fact I did sex education, or that I posted sex toys – it was a personal attack."

The identities of flaggers mentioned by participants seem to indeed vary by country. Marta, a high-profile Ukrainian creator, found Russian bots or accounts were the main players who tried to de-platform her, judging by the negative comments she received. UK-based transgender participants largely attributed mass reporting and de-platforming to Twitter mobs of so-called gender critical accounts, or those excluding transgender women from cisgender spaces and debates (Zanghellini, 2020). Italian users Ale, Elia and Gin referred to members of their hometown communities or to Telegram groups of middle-aged fascist sympathisers instead. Participants became aware of abusers' identities after they entered the groups, which regularly shared the Italian "quando c'era lui," or "when he" – 'he' being Mussolini – "was around." In other cases, users who had rejected advances from their followers believed their deletions were a direct result of this rejection, a personal campaign against them.

Geopolitics, social and cultural inequalities and nuances therefore showcase the global inequalities of platform governance, assembling with platform infrastructure against users' identities as much as to their content. This has meant that platforms have de-platformed sexuality or identity-focused content and posts by

marginalised groups, but not identical content posted by celebrities or hate against the same marginalised groups. Non-binary activist Ale, for example, described their experiences of being de-platformed after reclaiming homophobic slurs, showcasing different content moderation actions against marginalised communities as opposed to against dominant groups (Diaz & Hecht-Felella, 2021). Users, therefore, seem to think social media companies are selective about whose freedom of speech they prioritise: transphobic slurs, accusations of paedophilia or racism against people of colour has not been de-platformed in participants' experiences, while expressions of their gender identities, jokes about white people posted by white creators or sexual expression tend to be targeted by platforms' moderation.

Malicious flagging and de-platforming's offline impact on users and their circles

Consistently with previous studies on de-platforming and flagging (Author & Briggs, 2023; Myers West, 2018), users cited severe adverse emotions resulting from the assemblage of misused flagging and de-platforming, resulting in loss of work and inspiration to create content, and in serious stress on the back of what they understood as bullying from both platforms and users.

The work, community building and creativity that go into building oneself a platform have left users extremely distressed upon de-platforming. Consistently with Myers West (2018), being targeted by both platforms *and* their own followers or viewers left participants sharing feelings of low mood, stress, anxiety and fear. The uncertainty on the back of malicious flagging left users feeling insecure, fearful and distrusting of their own communities.

Users also found de-platforming and harassment to be an extension of the hate and exclusion they faced offline, in a space they once viewed as safe, creative and generative. This was particularly true for LGBTQIA+ and transgender users such as sex educator Gin and non-binary trans-feminine creator Bel:

- *"I created my profile to say, 'Enough with my past,' with my family who insisted I had to be female on their terms, find a husband, become a housewife. It was a rebellion, saying, 'This is me,' but once the profile goes, I felt that I couldn't be who I was anymore."*

- *"It was hurtful, emotionally damaging to think, 'Oh, no, I'm starting to love myself, but a lot of the world doesn't love who I am,' and then mixed with that, getting videos reported probably by the same people leaving those hate comments. So those comments but then also getting videos taken down when I'm just expressing my inner joy, it was just really stunting - it was a big blow to my self-confidence, because I was trying to accept myself and then on social media just being hit with a: 'No you can't.'"*

Even more concerningly, the harassment users faced resulted in increasing, never-ending spirals of abuse and de-platforming that continued throughout users' online and offline lives particularly against transgender participants, who were periodically targeted and even stalked and harassed to invalidate their earnings, their academic work, their book publishers and their earning prospects.

Additionally, participants shared that being the target of this form of platform-enabled and platform-perpetrated abuse had an impact not just on them directly, but

on their partners, friends, loved ones and communities too – a previously over-looked aspect in platform governance and de-platforming research, as showcased in these quotes by Elia and Gin:

- *"The deletion of [partner's profile on the back of flagging] has been the worst experience of my life so far. We already had a series of ongoing collaborations where we were waiting for payment, we'd just moved in together not even three months prior, we didn't even know what living alone meant and had no chance to ease ourselves in without this bomb dropping on us. I don't have a good relationship with my parents, so I didn't want to ask them for money, but the collaborations we had in place either didn't happen or didn't pay us because [partner's] profile was gone."*

- *"Not long before [our deletion on the back of misused flagging], a 13-year-old kid got in touch saying he was about to kill himself, that he couldn't take it any longer, so we spent a while night answering back, providing support and even finding charities and associations that could help him with the problem he raised with us, referring him to suicide helplines. […] What if this kid needed our help again? How would he have written? How would he not have killed himself? So emotionally, this was huge. It was a sense of physical loss."*

In this sense, participants' experiences seem to show that platform governance does not only affect them personally: it makes them unable to perform and fulfil the

duty of care that their following, their content has bestowed on them, showing the impact of de-platforming not just on individual users, but on wider communities.

Lastly, consistently with Author (2022), participants perceived receiving community guidelines violations, particularly for nudity, sexual activity and even soliciting, as a form of consent violation, as an interpretation of their content, activities and persona that they did not agree with and that they found upsetting. This consent violation assumed even sexist and transphobic dimensions: non-binary participants found their posts would stay up when they behaved and dressed in ways socialised as male, and would be removed when they were posing or dressing in ways socialised as female. For queer and trans people, algorithmic labelling was a denial of self-determination: "I don't think, 'Today I am female, today I am male' when I post pictures of myself, but the fact that the algorithm moderates me differently depending on which gender it thinks I am is worrying," said Ale, arguing that it is "transphobic that your choice of clothing, or the way that you're posing, determines how your algorithm understands you and defines you."

Conclusion

This study aimed to showcase how misused and malicious flagging manifests on users' profiles, and to highlight this behaviour's impact on users' online and offline lives.

Flagging is characterised by silence (Gerrard & Thornham, 2020), making its assemblage with de-platforming difficult to investigate without speaking to users: platforms share very little information about its uptake (Crawford & Gillespie, 2016), and users only see its result - i.e. de-platformed content or profiles, often

following online abuse or of virality (Author, 2022; 2023; Perrett, 2021; Stokel-Walker, 2022). This study therefore defines misused and malicious flagging on Instagram and TikTok as the misuse of platforms' flagging functionality to report content not because it violates social media community guidelines, but due to personal disagreement and/or attempts to harass targeted users to trigger their de-platforming. De-platformed users' experiences show users can ascertain that malicious and misused flagging is taking place because either those carrying out the reporting tell them they have flagged their content, or because contacts in and outside platforms alert them to coordinated, cross-platform abuse. However, often, users have no way to confirm they are being targeted with malicious flagging, particularly when this activity may come from single individuals misusing the flagging functionality outside collective action, on an individual basis. Therefore, following this study's findings, it is reasonable to believe that misused and malicious flagging is happening when a user's account is targeted with one or more of the following: receiving swathes of negative comments on viral content and/or on old posts, experiencing repeated content takedowns of new and old posts, removal of posting privileges or of the opportunity to go live, to like, comment or follow.

This study has found significant loopholes within Instagram and TikTok's flagging affordance, meaning that it can assemble with existing prejudice against specific content and profiles rooted in community guidelines and with users' agendas putting specific accounts and content particularly at risk of de-platforming via reporting. In particular, nuanced content such as nude or sex-related imagery and videos, LGBTQIA+ expression, journalism and activism seem to be prime misused or malicious flagging targets.

Participants shared the deep, troubling emotional and financial impacts the combination of abuse and de-platforming has on them, stunting their abilities to express and be their true selves, to reach their communities and networks and to be able to work – an impact that becomes greater offline, concerning their mental health, their safety and their ability to support themselves financially. As a result, it appears that platforms' over-reliance on automation to highlight dangerous content through flagging is facilitating, rather than preventing, online abuse. Yet, despite the crippling effects of this protection affordance turned abuse tool, users have found that nothing amongst platforms' reporting tools can help them notify Instagram and TikTok that malicious reporting is taking place. It is therefore recommended that platforms create ways for users to notify that malicious flagging is happening, and that they re-target the functionality to ask for more specific reasons behind reports to better identity malicious or misused activity, while also investing in contextual, human moderation to prevent the misuse of this tool. Regulators and platforms attempting to improve safety from online abuse and fairness when tackling of mistaken de-platforming should also pay particular attention to content and profiles that are already disproportionately affected by censorship, also allowing users to protest content moderation decisions by adding 'misused or malicious flagging' options to the appeals process, and adding a 'misused or malicious flagging' option to the existing avenues to report online abuse.

Bibliography

Author (2021) The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, 22(8):2002-2019.

Author (2022) An autoethnography of automated powerlessness: lacking platform affordances in Instagram and TikTok account deletions. *Media, Culture & Society,* 45(4): 822–840.

Author (2023) Flagging as a silencing tool: exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok. *New Media & Society,* forthcoming.

Author & Briggs, P (2023) The emotional and financial impact of de-platforming on creators at the margins. *Social Media + Society, 9*(1). Available at: https://doi.org/10.1177/20563051231155103 (accessed 7 June 2023).

Blunt, D and Wolf, A (2020) Erased: The impact of FOSTA-SESTA and the removal of Backpage on sex workers. *Antitraffickingreview.org Special Issue – Technology, Anti-Trafficking, and Speculative Futures*, 14: 117-121.

Braun, V & Clarke, V (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3: 77-101.

Braun, V & Clarke, V (2019) Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise & Health*, 11(4): 589–597.

Bronstein, C (2021) Deplatforming sexual speech in the age of FOSTA/ SESTA. *Porn Studies*, 8(4):367-380.

Caplan, R, & Gillespie, T (2020). Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society, 6(2)*. https://doi.org/10.1177/2056305120936636.

Castells, M (2001) *Internet Galaxy: reflections on the Internet, business, and society*. Oxford University Press.

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.* Available at: https://counterhate.com/research/hidden-hate/ (accessed 7 June 2023).

Clark-Flory, T (2019) A Troll's Alleged Attempt to Purge Porn Performers from Instagram. *Jezebel*. Available at: https://jezebel.com/a-trolls-alleged-attempt-to-purge-porn-performers-from-1833940198 (accessed 7 June 2023).

Coombes, E, Wolf, A, Blunt, D and Sparks, K (2022) Disabled Sex Workers' Fight for Digital Rights, Platform Accessibility, and Design Justice. *Disability Studies Quarterly*, 42 (2).

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.

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 (accessed 7 June 2023).

Cox, J (2021) Scammer Service Will Ban Anyone From Instagram for $60. *Vice*. Available at: https://www.vice.com/en/article/k78kmv/instagram-ban-restore-service-scam (accessed 7 June 2023).

Crawford, K and 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 Centre for Justice*. Available

at: https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf (accessed 7 June 2023).

Duffy, BE, & Meisner, C (2022) Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society*, 45(2): 285–304.

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*: 2371–2382. Available at: https://doi.org/10.1145/2858036.2858494 (accessed 7 June 2023).

Fiore-Silfvast, B (2012) User-generated warfare: a case of converging wartime information networks and coproductive regulation on YouTube. *International Journal of Communication*, 6:1-24.

García-Rapp, F (2017) Popularity markers on YouTube's attention economy: the case of Bubzbeauty. *Celebrity Studies*, 8(2):228-245.

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

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*, Special Issue on Media and Uncertainty, 16:1-19.

Goanta, C and Ortolani, P (2021) Unpacking Content Moderation: The Rise of Social Media Platforms as Online Civil Courts.' Forthcoming, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3969360 (accessed 7 June 2023).

Golbeck, J, Ashktorab, Z, Banjo, R, Berlinger, A, Bhagwan, S, Buntain, C, Cheakalos, P, Geller, A, Gergory, Q, Gnanasekaran, R, Gunasekaran. R, Hoffman, K, Hottle, J, Jienjitlert, V, Khare, S, Lau, R, Martindale, M, Naik, S, Nixon, H, Ramachandran, P, Rogers, K, Rogers, L, Sarin, M, Shahane, G, Thanki, J, Vengataraman, P, Wan, Z, Wu, D (2017) A Large Human-Labeled Corpus for Online Harassment Research.'*Proceedings of the 2017 ACM on Web Science Conference*: 229-233.

Google (n.d) Report inappropriate videos, channels, and other content on YouTube. *YouTube Help – Reporting and Enforcement*. Available at: https://support.google.com/youtube/answer/2802027/ (accessed 7 June 2023).

Graves, L (2007) The Affordances of Blogging, A Case Study in Culture and Technological Effects. *Journal of Communication Inquiry* 31(4): 331-346.

Hinduja, S and Patchin, JW (2007) Offline Consequences of Online Victimization. *Journal of School Violence*, 6(3):89-112.

Jane, E (2014) 'Back to the kitchen, cunt': speaking the unspeakable about online misogyny.' *Media & Cultural Studies* 28(4):558-570.

Keats Citron, D (2014) *Hate Crimes in Cyberspace*. Harvard University Press.

Leybold, M, & Nadegger, M (2023) Overcoming communicative separation for stigma reconstruction: How pole dancers fight content moderation on Instagram. *Organization*, 0(0). Available at: https://doi.org/10.1177/13505084221145635 (accessed 7 June 2023).

Malterud, K, Siersma, VD, Guassora, AD (2016) Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qualitative Health Research*, 26(13):1753-1760.

Mantilla, K (2015) *Gendertrolling - How Misogyny Went Viral*. Praeger.

Myers West, S (2018) Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms.' *new media & society*, 20(11): 4366–4383.

Navarro, R; Yubero, S and Larrañaga, E (2015). *Cyberbullying Across the Globe: Gender, Family, and Mental Health.* Springer.

Oversight Board (2023) Gender Identity and Nudity. *Decisions*. Available at: https://oversightboard.com/decision/BUN-IH313ZHJ/ (accessed 7 June 2023).

Paasonen, S, Jarrett, K and Light, B (2019) *#NSFW: Sex, Humor, And Risk In Social Media*. The MIT Press.

Perrett, C (2021) Transgender TikTok creators say the app's mysterious 'For You' page is a breeding ground for transphobia and targeted harassment. *Business Insider.* Available at: *https://www.businessinsider.com/tiktok-transphobia-problem-creators-report-harassment-threats-2021-2?r=US&IR=T* (accessed 7 June 2023).

Peterson, C (2013) User-Generated Censorship: Manipulating the Maps Of Social Media. *MIT Graduate Program in Comparative Media Studies*. Available at: https://cms.mit.edu/user-generated-censorship/ (accessed 7 June 2023).

Roberts, ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

Roulston, K (2022) *Interviewing: A Guide to Theory and Practice.* SAGE.

Schoenebeck, S and Blackwell, L (2021) Reimagining Social Media Governance: Harm, Accountability, and Repair. *Yale Journal of Law and Technology,* 23: 113-152.

Silverman, C & Fortis, B (2023). A Scammer Who Tricks Instagram Into Banning Influencers Has Never Been Identified. We May Have Found Him.

ProPublica. Available at: https://www.propublica.org/article/instagram-fraudster-ban-influencer-accounts (accessed 7 June 2023).

Simpson, J (2022). 'Whorephobia in Higher Education: a reflexive account of researching cis women's experiences of stripping while at university.' Higher Education, 84: 17–31.

Stokel-Walker, C (2022) TikTok Was Designed for War. *Wired*. Available at: https://www.wired.com/story/ukraine-russia-war-tiktok/ (accessed 7 June 2023).

Zanghellini, A (2020) Philosophical Problems With The Gender-Critical Feminist Argument Against Trans Inclusion. *SAGE Open*: 1-14.