Flagging as a silencing tool: exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok

Abstract

This study examines the unintended consequences of Instagram and TikTok's flagging functionality through a qualitative survey gathering the experiences of users who post content related to sex work, sexuality or nudity and who believed they were de-platformed following malicious reporting by other accounts. Using negative comments on de-platformed users' profiles as an indicator of malicious flagging and through a World Risk Society Framework, this paper explores the relationship between flagging and de-platforming, connecting the former's unintended outcomes with the opacity and inequalities of platforms' governance of nudity and sex. This study finds that, coupled with platforms' lack of transparency over content moderation decisions and with an inadequate appeals system, flagging can be misused in a vigilante fashion to silence creators of nude and sexual content due to the heightened censorship they face. This generates *more* risks for these users, who lose work and networks and become exposed to account recovery scams.

Keywords: flagging; content moderation; platform governance; censorship; Instagram; TikTok.

Introduction

Social media platforms are both crucial self-expression and public debate spaces (Gillespie, 2010) and a way for brands and professional content creators to strike up partnerships, sell products and promote themselves and their work (Glatt, 2022). However, the scale of content posted on social networks and the increased number of users who populate them means that they need to be heavily moderated to make sure they remain usable (Diaz and Hecht-Felella, 2021). The moderation, or the deletion or censorship, of online content is therefore a key aspect of platform governance which sees platforms decide what types of content are allowed and visible in their spaces. However, content moderation has often disproportionately targeted queer and marginalised users (Haimson et al., 2021), sex workers (Blunt et al., 2020), disabled and plus size users (Coombes et al., 2022; Joseph, 2019), triggering loss of earnings and affecting users' wellbeing and raising concerns around freedom of expression and governance inequalities (Author & Briggs, 2023).

Studies have linked content moderation's excessive focus on nudity and sexuality instead of on violence (Author, 2022; Duffy & Meisner, 2022; Lumsden & Morgan, 2017) to the 2018 Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) and the Stop Enabling Sex Traffickers Act (SESTA) (Bronstein, 2021; Haimson et al., 2021; Paasonen et al., 2019 etc). An exception to Section 230 of the United States' Telecommunications act intended to fight sex trafficking, FOSTA/SESTA caused platforms to over-censor posts worldwide for fear of being accused of enabling trafficking, applying US legislation to global content (Bronstein, 2021). Since then, sex workers' (SWs) accounts have been deleted without warning. Further, athletes, lingerie, sexual health brands, sex educators and activists have had their content deleted or shadowbanned, whereby "vaguely inappropriate content" is hidden from the platform's main pages without the user knowing (Author,

2021; 2022; Cotter, 2021). Creators posting content that platforms code as sexual are therefore often excluded from the visibility and work opportunities that platforms offer because their presence is viewed as inherently dangerous (Author & Paasonen, 2021).

As a mechanism allowing users to report content to social media platforms and express their concerns about their governance, flagging is both a crucial affordance to empower users (Crawford & Gillespie, 2016) *and* a functionality that can potentially be misused by malicious actors to trigger the aforementioned content governance techniques (Stardust et al., 2022). Indeed, flagging can also be weaponised against accounts other users disagree with, through "user-generated warfare (UGW)" (Fiore-Silfvast, 2012), making it a crippling online abuse technique aiding malicious actors in banishing users from online spaces, a silencing strategy against those who have been disproportionately targeted by content moderation, such as SWs (Stardust et al., 2022). And while nudity and sexuality have been bearing the brunt of platform censorship post-FOSTA/SESTA, the weaponization of this affordance by opposing factions in conflicts in Palestine and Ukraine have also triggered content removals, targeting activists and journalists alike (Stokel-Walker, 2022).

To understand the relationship between malicious flagging and de-platforming, this study focuses on two mainstream social media platforms used by professional content creators: the Meta-owned photo and video sharing platform Instagram, and its fast-growing Chinese rival, the video sharing platform TikTok, gathering user experience of their flagging functionality through a qualitative survey among accounts who post content related to sex work (SW), sexuality, or nudity in the US, United Kingdom, Europe and Australia and who believe they were de-platformed

following a report by other accounts. Focusing on negative comments as an indicator of malicious flagging and consequent de-platforming, this study connects flagging's unintended outcomes with the opacity and inequalities of platforms' governance of nudity, sex and SW. In doing so, it provides novel contributions to research by showcasing how flagging can be misused in a vigilante fashion due to the heightened censorship creators of nude and sexual content face, and highlighting the techniques users are forced to utilise when attempting to recover their deleted accounts.

Platform governance in the risk society

Platform governance is "a set of actions with wide-ranging socio-political effects on transnational communication and workers' rights" (Bronstein, 2021: 370). As part of this, content moderation is an integral part of the running of internet platforms (Goanta and Ortolani, 2021). Its rules are set by platforms' community guidelines or standards, which are often enforced by a blend of human and automated action (Gillespie et al., 2020).

Instagram and TikTok delete accounts that go against their community guidelines, often without warning (TikTok, n.d.; Instagram, n.d.). The Meta-owned platform's guidelines ban the support or praise of terrorism, organised crime or hate groups, and the buying or selling of firearms – serious harms somehow listed *together* with (and therefore likened to) the offering and selling of sexual services, aka commercial SW (Instagram, n.d.). The same guidelines ask users not to "spam people or post nudity," stating clearly that nudity, sex and SW are undesirable at best, dangerous at worst in the eyes of the platform. Nudity and sexuality are also

banned on TikTok, which prohibits content that is “overtly revealing of breasts, genitals, anus, or buttocks, or behaviours that mimic, imply, or display sex acts” (TikTok, n.d.).

Community guidelines are enforced through algorithms, "codified step-by-step processes" platforms use "to afford or restrict visibility" (Bishop, 2019: 2590). Algorithmically enforced community guidelines are similar across different platforms, covering content already prohibited by most countries’ criminal law (Gillespie, 2018; Goanta and Ortolani, 2021; etc). However, particularly post-FOSTA/SESTA, platforms have been notoriously anxious about hosting and over-zealous in moderating even legal nude and sexual content, reflecting social biases about the visibility of women’s bodies into online spaces, and showing nervousness about the use of said spaces for commercial sex (Ruberg, 2021). To this end, even some camming platforms have been rejecting any association with SW, limiting their users’ opportunities for labour and expression (Stegeman, 2021). Other social networks such as Instagram and TikTok (Author, 2022) have preferred largely banning nudity and sexuality altogether, leading to a progressive shrinking of online spaces for bodily and sexual expression (Author and Paasonen., 2021).

Instagram and TikTok’s categorising of something as humanly necessary as sex within the box of harms reveals their risk management strategy, which is connected to their commercial interests (Goanta and Spanakis, 2020), to modern anxieties surrounding technology (Hasan, 2018; Naak, 2010) and to their overzealousness in the face of FOSTA/SESTA (Blunt et al., 2020; Blunt & Wolf, 2020; Nolan Brown, 2022). Due to this emphasis on risk management, platform governance can then be conceptualised through Beck’s (1992; 2006) and Giddens’ (1998) World Risk Society theory.

Businesses and institutions in the World Risk Society address uncertainty by micromanaging risks produced by late modernity (Beck, 2006). While for Beck (1992) the driving forces of first modernity, or of the industrial society, were hunger and the distribution of wealth, the main driver of second or late modernity in our technologically advanced Risk Society is fear. More modernity and progress generate new risks brought by the same technologies that bring forth progress (Hasan, 2018). And while Beck's risks are often environmental (Beck, 1992; Hasan, 2018), the Risk Society's focus on technology-related risks in shared spaces can be applied to social media.

Risks are socially and ideologically mediated, and their consciousness often reflects the emotive consequences and meanings the recipients of risk messaging may attach to perceived risks rather than to their probability (Naak, 2010). This means that old fears, ideas and anxieties are reflected into modern society and its technologies, leaving individuals increasingly exposed to insecurity (Hasan, 2018). Because risks are socially constructed, the focus on preventing them means identifying certain populations as high-risk, further marginalising society's others (Beck, 2006).

The offline and online sex industry is a prime example of how identifying high-risk populations can other marginalised people. Mac and Smith (2019) argue that states' criminalisation of SWs, who tend to join the profession from already marginalised backgrounds due to hardship, race, mental or physical disability, gender identity or sexuality, homelessness, drug addiction etc., makes their lives worse, aiding instead of stopping violence from customers and the police alike. Although online SW is significantly safer, allowing SWs to vet clients, share information and advice on marketing, safety, rights and various work-related matters

(Sanders et al., 2020), banking services exclude SWs for fear that their transactions may result in fraud and chargebacks, while technology companies de-platform them due to risks grounded in North American, puritan perceptions of threats to business and reputation (Beebe, 2022; Paasonen et al., 2019; Stardust et al., 2020).

Because of the treatment faced by SWs, and, by extension, by those who create content coded by platforms as sexual (Author & Paasonen, 2021) in platform governance, World Risk Society theory can help analyse loopholes surrounding malicious flagging and censorship, particularly in the aftermath of FOSTA/SESTA. Already used to conceptualise platforms' use of shadowbanning of nude content, World Risk Society can be used to conceptualise the way platforms attempted to comply with FOSTA/SESTA (Author, 2021).

FOSTA/SESTA reflected old fears, ideas and biases that have so far negatively affected previously marginalised groups such as SWs and LGBTQIA+ people, and it has been dubbed as an expression of fears that technological development could increase sexual exploitation (Author, 2021; Haimson et al., 2021; Nolan Brown, 2022). Through the joint bill, the US government applied a World Risk Society logic to content moderation, forcing platforms to adopt its approach and transfer it not only to what was covered by FOSTA/SESTA, but also to adjacent content posted globally on platforms with large US audiences (Beebe, 2022; Blunt & Wolf, 2020; Nolan Brown, 2022). Big Tech have therefore fully embraced the World Risk Society, expanding it further: while FOSTA/SESTA has not been found to reduce sex trafficking and has only resulted in one charge so far (Blunt et al., 2020), its risk-focused governance has resulted in a significant chilling of online sexual expression, discussions around SW and pleasure, marketing of sex-positive content and brands, with policies "more restrictive than necessary to avoid liability," and platforms going

"into overdrive in their efforts to create distance from sexual speech and sex work to avoid potential legal problems and to attract investors and advertisers" (Bronstein, 2021: 368).

If businesses' and governments' increasing preoccupation with risks largely benefited private insurance firms, who defined and managed risks (Beck, 2006; Giddens, 1998; Hudson, 2003), platforms' governance of risk is inevitably connected not just to their audiences' anxieties, but also to their payment providers, their advertisers, their public image and their self-protection from legal charges. The risk platforms attempt to manage therefore is not only reputational: it is financial, tying governance of most online sex work and sexual expression to the US. Indeed, while laws governing sex work and sexual expression differ around the world, financial industry regulations are mostly US-based, meaning most nations must connect with US-based credit card companies and comply with their terms of service (Beebe, 2022). Therefore platforms, already largely born and based in the US and showing a puritanical approach to governance (Paasonen et al., 2019), must abide by the whims of the US' financial system even when, in the case of TikTok, they were not *born* or fully active in the US alone. This hits users with a double whammy of puritanical cultural and financial governance: platforms restrict nude and sexual content to shield themselves from the potential risks of platforming SW and sexual expression, and payment providers have been increasingly excluding businesses relying on SW from their services (Nolan Brown, 2022).

This risk-averse and puritanical financial and platform governance of sex shows financial discrimination is applied even to services that are legal in the US, such as the multi-million-dollar pornography industry or stripping, extending even to countries where governance of sex work is less strict (Beebe, 2022).

The affordances of flagging

FOSTA/SESTA led platforms to strengthen both content moderation actions such as shadowbanning and de-platforming to manage content visibility in their spaces (Author, 2021;2022) *and* affordances that outsource the responsibility of triggering said actions to audiences, such as flagging (Goanta and Ortolani, 2021). Originally developed in psychology (Gibson, 1977; Norman, 1988) and later used in communication studies (Bucher and Helmond, 2018; Graves, 2007), affordances can be viewed as a set of functionalities that allow both users *and* platforms to perform different actions. However, Nagy & Neff (2015) prefer using the concept of imagined affordances, since these are not merely shaped by technology and its designers, but also by users' interpretation of their use.

Similarly to liking, commenting and sharing, flagging is a feature platforms created for users to highlight content that potentially violates community guidelines (Crawford & Gillespie, 2016). Flags are afforded to both audiences *and* platforms themselves on Instagram and TikTok, and they are a mechanism for platforms to learn from users, distributing labour to volunteers (Crawford & Gillespie, 2016) and a form of crowdsourcing algorithmic progress "to facilitate the evasion of liability by platforms as intermediaries" (Goanta and Ortolani, 2021: 139). However, flagging's influence remains opaque: users have so far been largely unaware as to why or if their reports have been taken up, or indeed as to why their content has been reported (Crawford & Gillespie, 2016).

Yet, the user input provided by flagging makes sure that, like all affordances, the use of flags is shaped by users' communicative and practical agency, and by

their understanding of its function. As a result, flagging has already been weaponised against SWs by clients who had turned sour (Stardust et al., 2023), against LGBTQIA+ users (Oversight Board, 2023) and against disparate content audiences did not wish to see (Author, 2022). Examples of malicious flagging, also known as organized flagging (Crawford & Gillespie, 2016), user-generated censorship (Peterson, 2013) or UGW (Fiore-Silfvast, 2012), are many: conservative group Truth4Time's coordinated effort to flag pro-gay Facebook groups; countering online Cyber-Jihad YouTube videos; a coalition of incels joining forces to 'purge' adult performers from Instagram (Clark-Flory, 2019; Crawford & Gillespie, 2016; Fiore-Silfvast, 2012). On TikTok, too, malicious flagging has been used by pro-Russian accounts against Ukrainian activists (Stokel-Walker, 2022).

Case decisions and policy recommendations by Meta's independent oversight body, the Oversight Board (2023), have shown that platforms' over-reliance on automated governance can result in heightened reports, triggering de-platforming against LGBTQIA+ users. Flags can thus become an easily gamed way to police content for users who may view a gay kiss or an artistic nude as too inappropriate for social media (Crawford & Gillespie, 2016), turning into a form of "digilantism," or "politically motivated extrajudicial practices in online domains that are intended to punish or bring others to account in response to a perceived or actual lack of institutional remedies" (Jane, 2017: 461). This way, platforms like Instagram and TikTok can simultaneously *facilitate* online harms through affordances such as flagging and *perpetrate* them through their design and governance, by taking up malicious flags and de-platforming their targets (Schoenebeck and Blackwell, 2021).

Flags are also extremely convenient for platforms, who do not have to honour them, and who can use them to justify the removal of contentious content (Crawford

& Gillespie, 2016), meaning that companies' over-zealousness in adhering to legislation and common decency and its related incentivisation of quick responses to risks has facilitated the proliferation of malicious flagging (Griffin, 2021).

Flagging has left users feeling targeted not just by platforms' processes, but by the retaliation of audiences themselves (Duffy & Meisner, 2022; Myers West, 2018). Cox (2019; 2021) found that many de-platformed users paid hackers *working for* platforms to reinstate their accounts, showcasing the inadequacy and accessibility of appeals systems, and that malicious actors also pay hackers to *trigger* de-platforming, meaning malicious flagging can be considered a silencing strategy to drive users, and particularly women, LGBTQIA+ and marginalised users, off digital spaces. The impact of de-platforming on users, and particularly on creators, can be huge, since their career success (e.g. gaining brand partners or monetizing their audience) "is directly related to platformed indices of visibility (i.e. views, likes, favorites, shares)" (Duffy & Meisner, 2022: 2). There is thus scope to investigate flagging as a form of online abuse, similarly to online harassment, flaming, doxing and image-based abuse (Lumsden & Morgan, 2017).

In this increasingly uncertain governance scenario where particularly users who work, express themselves and communicate through their bodies find themselves under threat of de-platforming at the hands of both users and platforms with little or no options to appeal, their own explanations and interpretations of the governance they are targeted with become all we have to understand social media companies' opaque governance.

Resisting opaque platform governance with gossip

The opacity of platform governance means users are not always privy to what exactly triggered their account's deletion (Schoenebeck and Blackwell, 2021): although both Instagram and TikTok show users when their accounts accumulate multiple violations and may be at risk of deletion (Author, 2022), users do not always have access to specific information, e.g. whether deletions were caused by a single post, a succession of posts, one or a series of reports by other users. What is known following a decision to be reviewed by Meta's Oversight Board (2023) is that in the case of partial nudity picturing trans and non-binary bodies, as little as three reports were enough for content to be taken down by Instagram.

Faced with this lack of essential information users, and particularly professional content creators who rely on social media visibility to strike partnerships with brands and to earn money by selling products and/or promoting themselves (Glatt, 2022), have been recurring to 'algorithmic gossip,' or "communally and socially informed theories and strategies pertaining to recommender algorithms, shared and implemented to engender financial consistency and visibility on algorithmically structured social media platforms" (Bishop, 2019: 2602). While gossip is often dismissed as biased or frivolous, it becomes a knowledge resource for marginalised groups, a tool to fight power and facilitate resistance (Bishop, 2019). Similarly, Savolainen found that users fight platforms' algorithms through "algorithmic folklore," or "beliefs and narratives about moderation algorithms that are passed on informally and can exist in tension with official accounts" (Savolainen, 2022:2). Combining personal experience with platforms, media reports and friends' tales, user experience of flagging and its surrounding folklore can provide insights into the opaque functioning of algorithms (Bishop, 2019; Duffy & Meisner, 2022; Savolainen, 2022). Yet, while users' theories represent a form of resistance to platform

governance, they do not always result in successfully managing the algorithm: in fact, they are often denied by platforms themselves (Savolainen, 2022).

Platforms' outright denials of moderation targeted against specific communities, even in the face of their own apologies (Author, 2021), have been likened to gaslighting, a strategy used to both neutralise criticism and to undermine user experience (Cotter, 2021). This gaslighting, which Blunt et al. mention in the case of shadowbanning, can be applied to flagging, too, as users are both "made to feel crazy, as their reality is being denied publicly and repetitively by the platform" (Blunt et al., 2020: 79) and "incapable of assessing algorithms independently of what platforms say about them" (Cotter, 2021:14). Thus, in denying users' experiences and knowledge, platforms minimise their expertise which, given that a variety of women and LGBTQIA+ users have been affected by censorship *and* use gossip (Author, 2022; Author & Briggs, 2023; Bishop, 2019), shows a patriarchal, patronising approach to criticism.

Through this paper, I wish to instead harness the power of user gossip, approaching the study of algorithms through a feminist lens by broadening the sources we learn from and reframing who we think is an "algorithmic professional" (Bishop, 2019: 2602). By centring de-platformed users with experiences of malicious flagging in the study of potential links between reporting and censorship, this paper answers the following research questions:

RQ1: What is the relationship between negative comments and the removal of nude and sexual content and/or profiles on Instagram and TikTok?

RQ2: Which techniques do users utilise to recover their accounts?

Methods

This study focused on Instagram and TikTok as they are largely free to use mainstream social media platforms used for work, self-expression, organising and memory-making (Author, 2021; Duffy & Meisner, 2022).

Having sought and secured ethical approval from my institution, placing a specific emphasis on avoiding identifying participants at risk of over-moderation, data was gathered during a three-month period (June-August 2022) through an anonymous qualitative survey circulated through my own social media profiles: Facebook, Twitter, Instagram and TikTok. Qualitative surveys feature open-ended questions centred on a particular topic and presented in a fixed order to all participants (Braun et al. 2021). They can produce rich accounts of participants' experiences, allowing them to respond in their own words instead of having to select from multiple choice answers (Braun et al. 2021). Furthermore, given that the nature of this study meant engaging with marginalised communities with personal experiences of de-platforming resulting in psychological distress (Author & Briggs, 2023), qualitative surveys allowed them to present their own narrative, in their own safe space, without me interfering with their narration.

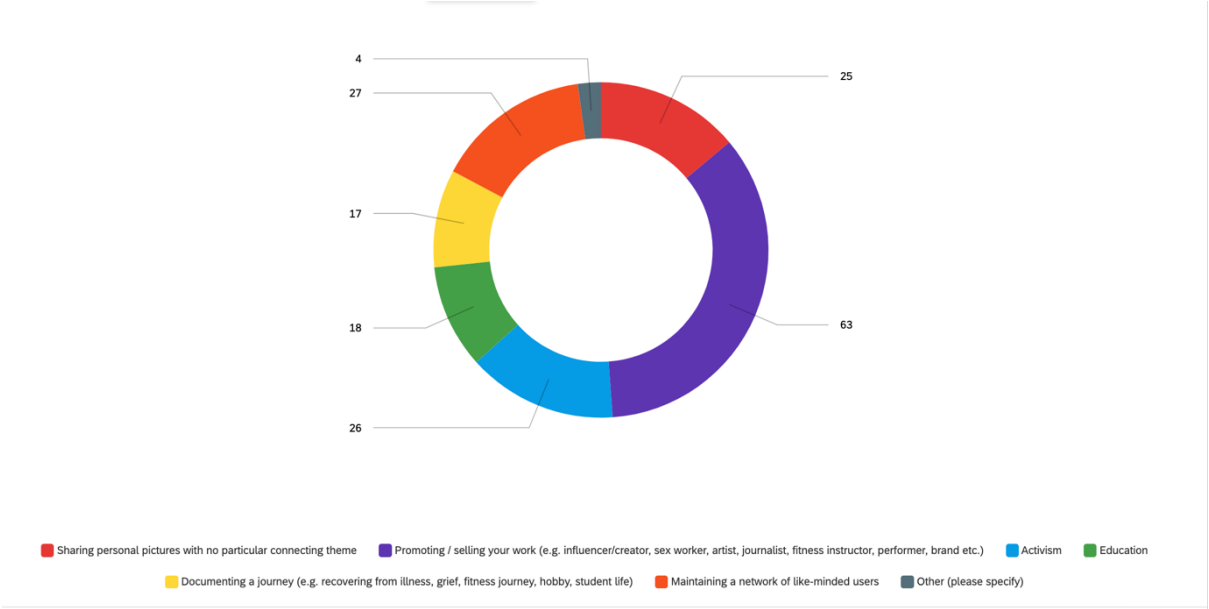
Inclusion criteria were as follows: participants had to be over 18 years of age and to be using social media to sell and/or promote themselves or their work. Crucially, they had to have experienced *both* negative comments *and* account and/or content deletion, a set of inclusion criteria which returned more specific experiences of platform governance post-FOSTA/SESTA and which mirrored my experiences of account deletion after receiving swathes of negative comments upon going viral

(Author, 2022) and other situations in which flags triggered de-platforming (Clark-Flory, 2019; Stokel-Walker, 2021;2022). As such, qualitative surveys are also relevant from "a social justice and inclusion point of view," as they "offer an accessible method to research beyond the 'usual suspects'" (Braun et al., 2021: 643;644).

I circulated this survey through my social media networks because, as a known researcher with experiences of de-platforming and of organising anti-censorship activism campaigns, I have a sizeable following of over 400,000, comprising artists, athletes, sex workers, activists, researchers and journalists with whom I have built relationships and who helped me share the call for participants. By circulating the survey amongst communities that knew me, I wanted to make sure users, and particularly sex workers, felt comfortable when sharing their experiences of de-platforming. My network, which included participants from various locations and different backgrounds or degrees of education, was prioritised over different avenues for data collection because of the experiences of suspected flagging they often raised in their posts. As such, participants were particularly interested in sharing these experiences, and did so even though no financial incentive was offered to take part in the survey. Additionally, XBIZ, a leading news website for the adult industry, wrote an article about this study, helping me circulate information about it amongst their audiences, which explains why SWs are over-represented in this paper (Parkman, 2022).

123 participants took part in this study. They were largely based in the UK, US, Australia and Europe, where my networks lie, and used both Instagram and TikTok almost interchangeably. 98 of them reported having been affected by both censorship *and* negative comments on Instagram and TikTok. 41 out of 123

participants provided full, detailed responses about their experiences. Users were asked the ways in which they used Instagram and TikTok, with most of them sharing they used platforms to promote and/or sell their work, in fields such as journalism, sex work, content creation, fitness etc. (n = 63). Participants also utilised platforms to keep in touch with like-minded users (n = 27), to engage in activism (n = 26), to share pictures without a connecting theme (n = 25), to share and receive education (n = 18), to document a journey (e.g. taking up a new sport, coming out as LGBTQIA+ etc.) (n = 17).



Most of the users surveyed were aged 25-34 (n = 18), with the second biggest group being 35-44 (n = 13) and the third 18-24 (n = 5). Respondents were largely white (n = 34), with only a few coming from mixed race (n = 3), Black (n = 2), Latine (n = 2) or Indigenous (n = 1) backgrounds. Most of those who took part in the survey were cisgender women (n = 28), followed by cisgender men (n = 5), non-binary (n = 3) and gender fluid (n = 2) users, with only one transgender woman taking part. Most respondents were bisexual (n = 13), followed by pansexual (n = 12), heterosexual (n

= 7), queer (n = 6) and gay (n = 2) users. Only 8 users identified as plus-size, and 16 identified as having a disability.

Survey questions were modelled on personal and documented experiences of de-platforming (Author, 2023; Stokel-Walker, 2021; 2022), asking participants 10 questions - a blend of open-ended and multiple choice, plus demographic-related questions for screening. The former featured questions about participants' experiences of de-platforming, their understanding of the reasons behind their deletion and the steps they took towards account recovery. Multiple choice questions were informed by previous preliminary findings showcasing marginalised users receiving negative comments were likely to be reported and, as a result, de-platformed (Author, 2022; Oversight Board, 2023; Stardust et al., 2023), and by research highlighting platforms' disproportionate censorship of BIPOC, LGBTQIA+, plus size and disabled users (Haimson et al., 2020). Towards screening of these protected categories, I therefore asked participants about their profiles' aims, their age, gender identity, sexual orientation, ethnic background, and whether they identified as plus size or as having disability, to screen for intersecting experiences of platform governance and background.

Responses are organised through thematic analysis (TA), via answer excerpts. TA allows researchers to identify, analyse and report themes or patterns within data (Braun & Clarke, 2006). Characterised by minimal data organisation, TA describes data sets in rich detail, providing insights into relevant lived experiences, particularly when researching on stigmatised and excluded communities (Braun & Clarke, 2006). Themes are developed through the creative labour of researchers' coding, and often reflect data collection questions, and their choice depends on whether they capture

relevant details about the research questions rather than on replicability (Braun & Clarke, 2019: 594).

This study presents a set of limitations. Firstly, conducting research on platforms *through* platforms was challenging: I experienced different levels of invisibility and, in some cases, an outright chilling effect on my ability to share research and collect data. While the call for participants received an exponentially reduced audience on my TikTok profile compared to my usual posts, Instagram flagged my survey – hosted on the conventional and university-approved survey platform Qualtrics – as a "dangerous link," potentially affecting the number of users who, already under threat of de-platforming and hacking, may have decided not to complete it. Due to the high shares my call for participants received amongst sex working communities, Instagram may have classed my link as spam or 'sexual solicitation' –- bitterly ironic, considering that platform governance is the topic of this study.

Secondly, the narrow inclusion criteria set during data collection meant that participants were limited to a specific experience, potentially excluding those who *may have been* maliciously flagged but who did not receive negative comments on their posts. Similarly, the sharing of the call for participation with my network, with its limited locations and demographics, addresses the very specific set of experiences of users based in the Global North. However, given users' previous experiences of malicious flagging (Clark-Flory, 2019; Fiore-Silfvast, 2021; Stokel-Walker, 2021; 2022), the inclusion criteria I set were the most tangible factor to continue investigating malicious reporting.

Finally, although some may wish to investigate flagging by speaking to flaggers themselves or to platforms, this study instead recognises the content moderation and

algorithmic expertise of those affected by it (Bishop, 2019), also due to little to no platform communications and transparency about governance processes (Author, 2023).

Analysis

Participants' experiences are presented via a thematic analysis, focusing on one crucial theme - user-perpetrated harm in the form of flagging (theme 1) – and its consequences, such as the platform-perpetrated harm of de-platforming, explained via algorithmic gossip (theme 2), and users' techniques to fight it via different means of account recovery (theme 3).

(1) Flagging as online abuse

Participants' experiences highlight significantly exploitable loopholes in platform governance, leaving particularly feminist, SWs and activist users vulnerable to silencing by those who disagree with them. Users highlight receiving negative comments on their posts to then be de-platformed shortly after, consistently with Author (2022; 2023) and Oversight Board (2023), as well as being flagged due to their political views or their references to being SWs in their social media posts or bios. Worryingly, users reported campaigns orchestrated by single individuals egging their followers on, even targeting sexual assault survival content.

"My thoughts are that I was maliciously reported by a fan that I had an online altercation with on my OnlyFans platform. He was an incel who had taken offence to

me being married and my Instagram was taken down on my wedding anniversary, after I posted about it, which I don't think was a coincidence."

"A person with a much larger following exploited their fans to abuse a system that is already broken. Once it hits a certain amount of reports it bans the content without any manual need or reason for appeal or approval."

Fully ascertaining that one or more reports caused de-platforming may be challenging without platforms' confirmation, but these users' experiences highlight that flagging can be an effective silencing strategy and online abuse technique by those who wish to target those running accounts that have already been disproportionately affected by platform censorship such as SWs, erotic art, body-positive accounts and so on (Author, 2021; Blunt & Wolf, 2020).

In this sense, an affordance such as flagging, created partly to mitigate risks on social media, is actually generating *new* risks, replicating offline inequalities against marginalised communities, who are othered in a repetition of the status quo (Beck, 1992; 2006; Hasan, 2018). Similarly to interpretations of World Risk Society then, platforms' moderation and their over-reliance on flagging and automation betrays their audiences' and their own anxieties in hosting nude and sexual content (Naak, 2010).

Malicious flagging exploiting these anxieties can then be likened to other deliberately harmful online behaviours such as flaming or doxing (Lumsden & Morgan, 2017). Harnessing Instagram and TikTok's track record of de-platforming sex, malicious flaggers can stage full-blown attacks (Fiore-Silfvast, 2012) against specific users, relying on over-zealous moderation processes and on time-poor and

under-paid moderators who have to make conservative decisions to deliver on their targets, leaving very little room for context (Suri and Gray, 2018). Therefore, online abusers are motivated to give flagging a shot, in the case that the very exploitable loophole of an under-funded, context-lacking platform governance can help them achieve their aims.

(2) Sex, nudity and SWs' vulnerability to de-platforming.

Following these perceived flagging and content moderation vulnerabilities, participants reported an array of content and profile deletions surrounding topics from SW to activism, from art to mental health advocacy, from memory-making to self-expression.

"My paintings have been removed several times without warning, […] flagged for nude content and sometimes sexual solicitation. Those featuring LGBTQ+ themes or POC tend to be removed faster and without hesitancy compared to my other pieces that depict white figures. Example: 2 paintings that had the same angle and pose were flagged, but the white model was kept on my page while the black model was removed."

"I was sharing LEGAL SAFE abortion pill information and it was deleted and my account is in jeopardy of being removed from instagram. No option to appeal either. Tiktok removed a pro-abortion meme of a historical painting of a woman bare breasted with a bunch of babies saying 'me meeting all of my abortions when I get to heaven.'"

"I have a business about destigmatizing STIs and fighting slut shaming. Essentially it's a place for empowering women. I've had a few posts taken down on Instagram but for reasons that didn't make sense. Like one time I shared info about abortion pills and it got taken down because it said I was selling illegal goods. I shared a personal story about sexual assault and my account was deleted within 24 hours."

Here, participants shared that intersecting elements of their identity, such as their work as SWs and their size or disability, their ethnicity and their sexual orientation, attracted the most negative comments ahead of de-platforming.

Consistently with research on the aftermath of FOSTA/SESTA, participants' experiences show that censorship of SW is trickling down to other groups (Author, 2021; Author and Paasonen, 2021). Sexual health and activism information were caught in the net of platform censorship. Participants' experiences show that bodily displays or women's health are a shorthand for sex on platforms, replicating previous research findings on content moderation sexualising bodies without users' consent (Author, 2022).

While some users claimed to have been deleted due to "plus size stigma," "misogyny" and due to "being a woman", SWs reported that merely associating with adult content platforms has been enough for them to be potentially flagged before being deleted by Instagram and TikTok. Although most users claimed to be following community guidelines, the mere mention to SW or OnlyFans in their biographies or link sections made them more vulnerable to deletion. Users recounted having had as many as six accounts deleted by Instagram and TikTok, despite showing no nudity:

"My content was taken down off of Instagram for 'nudity' when there was no nudity whatsoever. I showed too much of my shoulder. Other content was deleted for 'solicitation of sexual services' because I simply mentioned OnlyFans."

"I have had videos removed for adult nudity when I'm in a hoodie and sweatpants. […] Swimming with my kids has been nudity, skateboarding with my daughter was also somehow nudity. I've had my livestreams removed for solicitation when I'm just talking about being a single parent."

The above stories are further examples of the pitfalls within platforms' World Risk Society governance, with companies identifying the wrong risks while further marginalising and othering vulnerable groups, whose difference is perceived as deviance (Beck, 1992; 2006; Hudson, 2003). SW-related content is a case in point: platforms seem to be profiling users, and particularly sex workers, assuming that even when they are posting personal or non-sex work related content, they must be soliciting regardless. This assumption is reflected in the variety of users who claim to not even post nudity, but to only feature an OnlyFans link, and are de-platformed and excluded from the benefits of an online life.

This exclusion shows platforms greatly underestimate their roles in their users' lives: far from being just a creative outlet or a tool to promote work, TikTok and Instagram are also a portfolio, a way to access vital information, to network with one's community and to make memories with friends and family (Author & Briggs, 2023). In assuming that users may be soliciting 24/7, Instagram and TikTok are profiling SWs and excluding them from the work, networking and information opportunities their platforms provide, stigmatising their work and denying them any

other sort of human interaction. This directly endangers SWs and marginalised communities, given that the loss of access to networks, funds and safety information has been found to have crippling financial and emotional consequences (Author & Briggs, 2023) and that accessing online spaces is crucial to SWs' safety (Sanders et al., 2020).

(3) Battling faceless governance

A striking majority of respondents were not successful at recovering their Instagram and/or TikTok accounts through platforms' official appeal systems. Consistently with the Oversight Board's findings (2023), this highlights significant failures within Instagram and TikTok's moderation infrastructure and an over-reliance on flags and automated moderation that results in excessive censorship and leaves users feel lost and disheartened.

Participants reported a lack of clarity, accessibility and transparency by platforms when dealing with account deletions. Often, despite self-censoring and following community guidelines, they were deleted regardless, and were not able to speak to platform employees to ask questions. Responses to my questions surrounding the steps users had taken to recover their accounts were a succession of: "I tried contacting Instagram / TikTok and never heard back." When dealing with the distressing aftermath of account deletions that led to loss of income and of network, most users reported not being able to reach human moderators or platform workers and having to deal with slow or ineffective appeals systems, even receiving wrong appeal links from platforms, or facing deletion despite having been a victim of hacking.

“I’ve had countless days of appealing to the faceless and nameless masters of Instagram. […] I scrounged for any direct connection; emailing people who never answered me, trying any way of reaching to someone who’d understand that this is my livelihood.”

“I tried for over two months to get my page back only to hit a brick wall. I don’t think I ever spoke to a human ever.”

Without being able to reach the “nameless masters” of platforms or “to speak to a human ever,” most participants could not understand *which* specific element of their content triggered their deletions, meaning that, were they to create other accounts, they may once again be facing censorship due to the inability to learn from previous experiences. Further, with respondents finding their appeals were not processed or reviewed, they were left in limbo: without hearing back from platforms, they had to take matters into their own hands. In these cases, users had two possibilities: striking a deal with hackers to attempt to retrieve their accounts or resigning themselves to the fact that their profiles will not be coming back.

With most participants reporting it is shared knowledge that platforms do not engage with deleted users, many have given up on recovering their banned accounts entirely. Users therefore had to start over, creating new accounts that may still be under threat of de-platforming, while also having to engage in the taxing digital labour of re-building a following. In line with Cox’s findings (2019; 2021), platforms’ lacking appeals systems made it impossible for participants to recover their accounts through conventional means. The possibility to lose their work and

networks therefore led them to spend money and run risks in the hope to recover their profiles:

"I have emailed them every day for weeks with no reply. I even paid a guy to get my instagram account back and he eventually gave up and just took my money and I never got my account back. I have had friends make reports for me and they never got replies back."

"[My deletion] also lent more legitimacy to scammers who use my content to create catfish profiles. If those stay up while my genuine profile is deleted then people are unlikely to see my real profiles with lower follow numbers as genuine."

Social media platforms' approach to moderation has therefore inadvertently created a market for scammers, from hackers promising to recover accounts (Cox, 2019) and to delete accounts (Author, 2023; Cox, 2021), all the way to those posing as a deleted profile to scam others out of nude images or money (Author, 2018).

Ironically, then, platforms' management of risk through new technologies has only generated *more* risks for users, who find themselves financially and emotionally lost after de-platforming (Author & Briggs, 2023) and have to resort to illegal means as a result.

Participants also reported that their communities have had to step in and fill the gaps left by platforms' inaction through awareness-raising, pulling contacts' strings and gathering support from freedom of expression focused non-profits. Users' followers reported them as 'missing,' submitting Help Centre tickets and emailing platforms:

"When my first account was banned, I was able to receive assistance from dozens of friends, and hundreds of followers, who all messaged/emailed tiktok to request I get my account back. Sadly they ignored all of these and never restored my account, which is why I created my new one."

"I have asked people that follow me to help with complaints and messages. I also contacted people inside the platform."

Consistently with previous experiences of de-platforming following suspected flagging (Author, 2022; 2023; Stokel-Walker, 2021), participants felt more supported by their own communities and often recovered their accounts only when someone in their network was able to directly speak to platform insiders. This highlights a shared understanding among marginalised communities that platforms' infrastructure is at best inefficient and inaccessible, and at worst directly biased against them. Therefore, rather than presiding over more efficient and welcoming spaces, new technologies seem to have generated more risks, recreating old power structures (Beck, 1992; 2006; Naak, 2010).

Conclusion

This paper highlighted the exploitable loopholes Instagram and TikTok's flagging affordances provide to malicious actors, increasing the vulnerabilities of those already under threat of de-platforming due to their work, their characteristics and/or the content they post. The experiences shared seem to show that Instagram and

TikTok allowed personal or moral crusades against them, affording power over content and profiles to anyone but the users who create them.

The joint power of malicious flagging and platforms' overzealousness in de-platforming sex resulted in the following isolating experiences: trying to recover one's profile, rebuilding a network and a workplace from scratch, having to rely only on their network, on scammers, on individuals with contacts within Instagram and TikTok to still be able to work or, otherwise, having to resign themselves to the loss of their profile and subsequent loss of livelihoods, memories and networks.

Participants felt they had been targeted with flagging, negative comments *and* de-platforming due to a combination of their content and their identities, as flaggers relied on platforms' aversion to nudity, sex and SW to silence them (Oversight Board, 2023; Stardust et al., 2023). While participants' beliefs about the moderation they received might not coincide with Instagram and TikTok's actual agendas, in line with Bishop (2019), I take their experiences seriously. As experts in their fields, about their communities and their subcultures, these users are highlighting failures within platform governance which are leaving them vulnerable, contrary to Instagram and TikTok's argument that it is aimed at users' safety (Instagram, n.d.; TikTok, n.d.), Users' experiences show that platforms are both *facilitating* and *perpetrating* harms against them (Schoenebeck & Blackwell, 2021), allowing others to report them and triggering their de-platforming. Their experiences are both valid and concerning.

My findings show that, similarly to other online abuse techniques such as flaming or doxing (Lumsden & Morgan, 2017), malicious flagging is a silencing strategy driving users offline, a strategy particularly effective against users and topics that are stigmatised or have been the target of platform governance: SW, art, activism, queer expression and sexual health education. Intentionally or not, those

who engage in flagging are playing a part in users' loss of network, livelihood and education, exploiting a platform governance that is already unequal (see Author, 2021; Blunt & Wolf, 2020; Haimson et al., 2020; Stardust et al., 2023) and skewed against marginalised communities, and that thrives off of opacity (Crawford & Gillespie, 2016).

What is clear from this paper's findings is that SWs have become increasingly vulnerable to this type of behaviour post-FOSTA/SESTA (see Blunt & Wolf, 2020; Stardust et al., 2023), and that platform governance seems to be blending with malicious flaggers' moral code, bringing offline whorephobia, or the hatred and disgust towards sex workers, into an already whorephobic, risk-focused platform governance (Beebe, 2022; Blunt & et al., 2020). The intersecting identities of those engaging with SW, nude and sexual content creation and sexual expression are further aggravating factor of malicious flagging: targets among participants included users with disabilities, low incomes and/or childcare needs who had finally found types of online work and expression that fit with their living situation and needs questions the supposedly safety-focused governance of content and profiles on Instagram and TikTok. This begs the question: *safety for whom?* Clearly, not for SWs and those engaging in sexual or nude expression, or those relying on platforms to make a living.

Malicious flagging's success rests on the fact that platforms do very little to restore accounts they disable, showing a carceral, punitive form of governance that does not allow for rehabilitation (Schoenebeck and Blackwell, 2021). The experiences shared by participants in this study therefore show a trigger-heavy approach to governance, a facilitation of flagging loopholes, an inadequate appeals system and an approach to user safety which is cavalier at best, and careless at

worst. Through withholding information about their practices and processes and through over-policing sexual content, social media are effectively handing the reins of their most conservative tools to misogynist, abusive accounts who benefit from the de-platforming of sex, sex work, nudity and content surrounding these topics.

Technology has considerable opportunities to de-stigmatise sex, to create safe spaces for marginalised users to work safely, organise, learn and share information. Instead, mainstream social media platforms such as Instagram and TikTok, on which a variety of users rely for work and connection, are governing their spaces through a World Risk Society approach, replicating offline inequalities through technology, and even generating further risks for marginalised communities. Platforms should instead work towards providing more clarity about their decisions to de-platformed users, specifying when content has been flagged by others and allowing accounts more room to appeal and contest those decisions. They should also increase transparency about the inner workings of their algorithms and improve their communications with and due process towards de-platformed users, to avoid the proliferation of scams to restore content and profiles.

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.

Bibliography

Author (2018) Does Instagram's verification discriminate against nudity? *Blogger On Pole.* Available at: *https://bloggeronpole.com/2020/08/does-instagrams-verification-process-discriminate-against-nudity/* (accessed 8 January 2024).

Author (2021) The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies* 22(8): 1-18. https://www.tandfonline.com/doi/full/10.1080/14680777.2021.1928259.

Author (2022) An autoethnography of automated powerlessness: lacking platform affordances in Instagram and TikTok account deletions. Media, Culture & Society, 45(4), 822-840. https://doi.org/10.1177/01634437221140531.

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

Author and 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

Author and Paasonen S (2021) Sex in the Shadows of Celebrity. *Porn Studies* 8(4): 411-419. https://www.tandfonline.com/doi/full/10.1080/23268743.2021.1974311.

Beck U (1992) *Risk Society: Towards a New Modernity*. Translated by Mark Ritter. London: Sage.

Beck U (2006) Living in the World Risk Society. *Economy and Society* 35 (3): 329–345.

Beebe B (2022) 'Shut Up and Take My Money!': Revenue Chokepoints, Platform Governance, and Sex Workers' Financial Exclusion.' *International Journal of Gender, Sexuality and Law*, 2: 140-170.

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

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.

Blunt D, Coombes E, Mullin S and Wolf A (2020) Posting Into the Void. *Hacking/Hustling*. Available at: https://hackinghustling.org/posting-into-the-void-content-moderation/ (accessed 8 January 2024).

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

Braun V, Clarke V, Boulton E, Davey L & McEvoy C (2021) The online survey as a qualitative research tool. *International Journal of Social Research Methodology*, 24(6): 641-654.

Braun V & Clarke V (2019) Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise & Health*, 11(4): 589–597. https://doi.org/10.1080/2159676X.2019.1628806.

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

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

Bucher T and Helmond A (2018) The affordances of social media platforms. In: Burgess J, Poell T and Marwick A (eds) *The SAGE Handbook of Social Media*. London and New York: Sage, pp. 233–253.

Clark-Flory T (2019) A Troll's Alleged Attempt to Purge Porn Performers from Instagram. *Jezebel*. https://jezebel.com/a-trolls-alleged-attempt-to-purge-porn-performers-from-1833940198 (accessed 8 January 2024).

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. https://www.tandfonline.com/doi/full/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 (accessed 8 January 2024).

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 8 January 2024).

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 and Hecht-Felella L (2021) Double Standards in Social Media Content Moderation. *Brennan Centre For Justice*, pp. 1-39.

Duffy BE & Meisner C (2022) Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society* 5(3.4): 103-107. https://doi.org/10.1177/2057047320959855.

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.

Gibson JJ (1977) The theory of affordances. In: Shaw R & Bransford J (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Erlbaum.

Gillespie T (2010) The Politics of Platforms. *New Media & Society,* 12 (3): 347–364. doi:10.1177/1461444809342738.

Gillespie T; Aufderheide P; Carmi E; Gerrard Y; Gorwa R; Matamoros-Fernández A; Roberts ST; Sinnreich A; Myers West S (2020) Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4). https://doi.org/10.14763/2020.4.1512.

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. https://ijoc.org/index.php/ijoc/article/view/15761.

Goanta C and Spanakis J (2020) Influencers and Social Media Recommender Systems: Unfair Commercial Practices in EU and US Law. *TTLF Working Papers* (54): 1-27.

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

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

Griffin R (2022) The Sanitised Platform. *JIPITEC*, 13 (36): 36-52.

Haimson O; Delmonaco D; Nie P and Wegner A (2021) Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction*,5, (CSCW2), No.: 466:1–35, https://doi.org/10.1145/3479610.

Hasan MN (2018) Techno-environmental risks and ecological modernisation in "double-risk" societies: reconceptualising Ulrich Beck's risk society thesis. *The International Journal of Justice and Sustainability*, 23: 258-275. https://doi.org/10.1080/13549839.2017.1413541.

Hudson B (2003) *Justice in the Risk Society: Challenging and Re-affirming Justice in Late Modernity*. London: SAGE.

Instagram. n.d.c. https://help.instagram.com/477434105621119 (accessed 8 January 2024).

Jane EA (2017) 'Dude … stop the spread': Antagonism, agonism, and #manspreading on social media. *International Journal of Cultural Studies*, 20(5): 459–475.

Joseph, C. (2019) 'Instagram's murky 'shadow bans' just serve to censor marginalised communities.' *The Guardian*. Available at: https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive (accessed 8 January 2024).

Lumsden K and Morgan H (2017) Media framing of trolling and online abuse, silencing strategies, symbolic violence and victim blaming. *Feminist Media Studies*, 17 (6): 926-940.

Mac J. & Smith M. (2018) *Revolting prostitutes: the fight for sex workers' rights*. London: Verso Books.

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

Naak SJ (2010) Contextualising risk, constructing choice: Breastfeeding and good mothering in risk society. *Health, Risk & Society*, 12 (4): 345-355. https://doi.org/10.1080/13698571003789666.

Nagy P & Neff G (2015). Imagined Affordance: Reconstructing a Keyword for Communication Theory. *Social Media + Society*, 1(2). https://doi.org/10.1177/2056305115603385.

Nolan Brown E (2022) The New Campaign for a Sex-Free Internet. *Reason*. Available at: https://reason.com/2022/04/09/the-new-campaign-for-a-sex-free-internet/ (accessed 8 January 2024).

Norman DA (1988) *The Psychology of Everyday Things*. New York: Basic Books.

Oversight Board (2023) Gender identity and nudity. *Decisions*. https://oversightboard.com/decision/BUN-IH313ZHJ/ (accessed 8 January 2024).

Parkman D (2022) Sex Workers, Performers Invited to Participate in Platform Discrimination Study. *XBIZ*. https://www.xbiz.com/news/266649/sex-workers-performers-invited-to-participate-in-platform-discrimination-study (accessed 8 January 2024).

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 8 January 2024).

Ruberg B (2021) 'Obscene, pornographic, or otherwise objectionable': Biased definitions of sexual content in video game live streaming. *New Media & Society*, 23(6): 1681–1699. https://doi.org/10.1177/1461444820920759.

Sanders T; Scoular J; Campbell R; Pitcher J; Cunningham S (2020) *Beyond the Gaze: Summary Briefing on Internet Sex Work.* https://www.beyond-the-gaze.com/wp-content/uploads/2018/01/BtGbriefingsummaryoverview.pdf.

Savolainen L (2022) The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society*, 4(6) 1–19. https://journals.sagepub.com/doi/10.1177/01634437221077174.

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

Stardust Z; Blunt D; Garcia G; Lee L; D'Adamo K & Kuo R (2023). High Risk Hustling: Payment Processors, Sexual Proxies, and Discrimination by Design. *City University of New York Law Review,* 26(1): 57-128.

Stardust Z, Garcia G and Egwatu C (2020) What can tech learn from sex workers? Sexual Ethics, Tech Design & Decoding Stigma. *Berkman Klein Center Collection*. https://medium.com/berkman-klein-center/what-can-tech-learn-from-sex-workers-8e0100f0b4b9 (accessed 8 January 2024).

Stegeman HM (2021). Regulating and representing camming: Strict limits on acceptable content on webcam sex platforms. *New Media & Society*, *0*(0). https://doi.org/10.1177/14614448211059117.

Stokel-Walker C (2021) TikTok censored a pole-dancing PhD who studies how social media silences women. *Input Mag.* https://www.inputmag.com/culture/tiktok-censored-banned-pole-dancer-phd-authorname (accessed 8 January 2024).

Stokel-Walker C (2022) TikTok Was Designed for War. *Wired*. https://www.wired.com/story/ukraine-russia-war-tiktok/.

TikTok (n.d) Adult nudity and sexual activities. *Community Guidelines*. https://www.tiktok.com/community-guidelines?lang=en#30 (accessed 8 January 2024).