Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

SEXUAL CONTENT MODERATION

Alexander Monea
George Mason University

Carolina Are
Northumbria University

Hanne Stegeman
University of Amsterdam

Rébecca Franco
University of Amsterdam

Zahra Stardust
Queensland University of Technology

Despite popular understandings of the internet as both teeming with pornography and offering safe harbor for LGBTQIA+ content, recent scholarship has documented the myriad ways in which internet platforms are cracking down on sexual expression online (Are 2022, 2023a, 2023b, 2024; Are & Briggs, 2023; Beebe 2022; Blunt & Stardust 2021; Gehl, Moyer-Horner, & Yeo 2017; Griffin 2024; Monea 2022, 2023; Myles, Duguay & Echaiz 2023; Stegeman 2021; Williams 2023). Content moderation is a core component of any internet platform (Gillespie 2018) and as part of their content moderation efforts most internet platforms have a series of one-size-fits-all, often puritan, adult and sexual content policies and content moderation practices for digital sexual expression and work (Monea 2022). In tracking the history of the hashtag 'NSFW' (Not Safe for Work), Susanna Paasonen, Keith Jarrett, and Ben Light (2019) show how contemporary platform discourse increasingly articulates sex and sexuality as "inherently risky, potentially harmful, and best hidden away and left unmentioned," (pp. 9-11) and ignores how sexual content and communication online is "key to people's self-definitions, central in terms of their wellbeing, and elementary in the building of social connections" (pp. 48-49). Deplatforming sex and sexuality has very real and material impacts on users, ranging from sex workers losing the capacity to safely solicit, screen, and meet with clients (Blunt & Stardust 2021; Monea 2022) to LGBTQIA+ users having their identities censured and their communities destroyed (Monea 2022;

Suggested Citation (APA): Monea, A., Are, C., Stegeman, H., Franco, R., Stardust, Z. (2024, October). *Sexual Content Moderation*. Panel presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

Tiidenberg, Henry & Abidin 2021). These impacts are not distributed equally and tend to disproportionately increase the precarity of those who are disabled, fat, Black, working class, or who are targeted for social marginalization due to other identity traits (Are & Briggs 2023; Monea 2022; Paasonen, Jarrett, & Light 2019).

This panel continues the ongoing research documenting the censorship of sexual expression online and extends it in new directions by examining its impact on distinct subcommunities. These analyses push current research in new directions by adding new findings and concepts to our understanding of the impact of deplatforming, shadowbanning, and other erasures of sexual expression online. Carolina Are, for example, examines changes in the shadowbanning of the pole dancing community on Instagram following the platform's apparent shift towards transparency, since they started enabling users to appeal a 'non-recommendable' account status. Pole dancers constitute a unique case study because their art form has roots in sex work but the art form itself is not necessarily sexual or performed by sex workers. This adjacency renders them vulnerable to overbroad content moderation practices. Author 1 draws on qualitative survey data from 100 pole dancers using Instagram, as well as ethnographic and autoethnographic research, showing how the new appeals are a cosmetic change to a faulty system. By documenting how shadowbanned pole dancers navigate Instagram's post-2023 shadowban appeals process, Author 1 shows the impact that trying to maintain their audience and continue to create regular content in a creatively stunting environment has on their everyday lives, both online and off.

Similarly, Alexander Monea documents cisnormative and heteronormative content moderation practices at TikTok that were especially prominent during its early years and examines the impact that this has had on 'Queer TikTok'. Despite TikTok being recognized as the most queer friendly social media platform today, creators of queer TikTok content still engage in self-censorship and heavily curtail their content to match their understanding of the opaque content moderation practices on the platform. This is shown through an analysis of TikTok's community guidelines and related policies over time, archival analysis of instances of biased content moderation, autoethnography of participation in Queer TikTok, and ongoing ethnographic interviews with Queer TikTokers. Drawing on this, Author 2 examines the ways Queer TikTokers have collectively produced 'folk remedies' for how best to evade censorship on the platform and argues that the continued prominence of these folk remedies on Queer TikTok demonstrates the lasting impact that bad content moderation can have on the algorithmic imaginary of a platform's user base.

Hanne Stegeman and Rébecca Franco turn to erotic webcam streamers to examine the ways in which sexual content moderation not only polices the permissibility of their content across internet platform, but also works to manage their labor on internet platforms. Based on this examination, they argue more (legislative) attention must be paid to moderation as labor management. The authors analyzed platform policies, conducted interviews with industry experts, attended adult industry conferences, and conducted interviews with 67 webcam performers. They show how sexual content moderation is frequently leveraged by internet platforms to discipline and control the labor of sex workers, with a specific focus on regulating their labor to direct clients across platforms and ensure that revenues remain funneled towards their platform.

Zahra Stardust examines how new and emerging independent small-scale platforms and cooperatives moderate sexual content. These platforms and their users often are explicitly opposed to the discriminatory, sex-negative, and surveillance capitalism based business models of major internet platforms and governments, and thus intentionally craft spaces that are meant to be more ethical and equitable. Author 5 conducted 10 hour-long qualitative interviews with new or emerging platforms. They document the challenges of reimagining internet platforms and content moderation, some of the most forward-looking and equitable ways of structuring internet platforms and content moderation, and the incredible difficulty of maintaining any alternative business model in the current political and economic landscape.

References

Are. C. (2024). Flagging as a Silencing Tool: Exploring the Relationship Between De-Platforming of Sex and Online Abuse on Instagram and TikTok. *New Media & Society, 0*(0). https://doi.org/10.1177/14614448241228544

Are, C. (2023a). 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

Are, C. (2023b). The Assemblages of Flagging and De-Platforming Against Marginalised Content Creators. *Convergence, 0*(0). https://doi.org/10.1177/13548565231218629\

Are, C. (2022). The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram. *Feminist Media Studies, 22*(8), 2002-2019.

Are, C., & Briggs, P. (2023). The Emotional and Financial Impact of De-Platforming on Creators at the Margins. *Social Media + Society, 9*(1). https://doi.org/10.1177/20563051231155103

Beebe, B. (2022). “Shut Up and Take My Money!”: Revenue Chokepoints, Platform Governance, and Sex Workers’ Financial Exclusion. *International Journal of Gender, Sexuality & Law, 2*(1-2), 140-170.

Blunt, D., & Stardust, Z. (2021). Automating Whorephobia: Sex, Technology and the Violence of Deplatforming: An Interview with Hacking//Hustling. *Porn Studies*, *8*(4), 350-366.

Gehl, R. W., Moyer-Horner, L., & Yeo, S. K. (2017). Training Computers to See Internet Pornography: Gender and Sexual Discrimination in Computer Vision Science. *Television & New Media*, *18*(6), 529-547.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. New Haven, CT: Yale University Press.

Griffin, R. (2024). The Heteronormative Male Gaze: Experiences of Sexual Content Moderation among Queer Instagram Users in Berlin. *International Journal of Communication, 18*(2024), 1266-1288.

Monea, A. (2023). I Know It When I See It: The Heteronormativity of Google's SafeSearch. *Porn Studies, 10*(2), 135-153. https://doi.org/10.1080/23268743.2022.2086163

Monea, A. (2022). *The Digital Closet: How the Internet Became Straight*. Cambridge, MA: MIT Press.

Myles, D., Duguay, S., Echaiz, L.F. (2023). Mapping the Social Implications of Platform Algorithms for LGBTQ+ Communities. *Journal of Digital Social Research, 5*(4), 1-30. https://doi.org/10.33621/jdsr.v5i4.162

Paasonen, S., Jarrett, K., & Light, B. (2019). *NSFW: Sex, Humor, and Risk in Social Media*. Cambridge, MA: MIT Press.

Stegeman, H. M. (2024). Regulating and representing camming: Strict limits on acceptable content on webcam sex platforms. New Media & Society, 26(1), 329-345. https://doi.org/10.1177/14614448211059117

Tiidenberg, K., Hendry, N.A., & Abidin, C. (2021). *Tumblr*. New York, NY: Wiley.
Williams, J. (2023). Deplatforming Sex Education on Meta: Sex, Power, and Content Moderation. *Media International Australia, 0*(0). https://doi.org/10.1177/1329878X231210612

PAPER 1: "SEXY JAIL": POLE DANCERS' DIGITAL LABOUR AND RESISTANCE STRATEGIES TO INSTAGRAM'S "ACCOUNT STATUS" UPDATE TO SHADOWBANNING

Carolina Are
Northumbria University

'Sexy jail' is a user-generated term online communities of pole dancers have begun utilising to describe Instagram's shadowban 2.0, or the notification their profile is 'non-recommendable' to other accounts according to the platform's algorithm due to the potential violation of nudity and sexual activity recommendations guidelines. This paper evaluates their experiences of posting during, navigating and appealing the platform's updated shadowban through a platform governance and digital labour perspective, in order to provide recommendations to users and social networks alike about the fairness, transparency and effectiveness of their tools.

Shadowbanning is a cross-platform light yet insidious social media censorship technique used by the likes of Twitter, Instagram and TikTok (Are, 2021a). Although social media companies do not use the term, it has become shorthand for their demotion of content and profiles, which are hidden from or not recommended to their main discovery feeds, greatly affecting the visibility, earnings, wellbeing and communications of the users affected (Are, 2021a; Blunt et al., 2020; Cotter, 2021). Known to particularly affect those who conduct work through social media - e.g. content creators, small businesses and marginalised communities such as LGBTQIA+ and BIPOC accounts (ibid; Duffy and Meisner, 2022) - shadowbanning is a controversial aspect of platform governance in continuous development.

Although not as damaging as outright de-platforming, or content and account removal from a social media platform, shadowbanning can greatly affect those whose work depends on visibility, such as content creators (Glatt, 2022), resulting in the time-consuming sharing and testing of strategies to game the algorithm known as 'algorithmic gossip' (Bishop, 2019).

Pole dancers have been at the forefront of protest against and knowledge sharing about shadowbanning, having been some of the first 'mainstream' users affected (Leybold & Nadegger, 2023). Practicing an art and a sport that has roots in the sex industry, not all pole dancers are sex workers but they create sex work adjacent content, making their posts a liminal space that requires context and that confuses algorithms and platforms' policy-makers alike (Are & Paasonen, 2021). In the aftermath of FOSTA/SESTA, the 2018 exception to Section 230 of the US Telecommunications Act 1996 that made social media companies legally liable for facilitating sex trafficking (a crime) and sex work (a job), leading them to over-censor swathes of content to avoid liability (Blunt & Wolf, 2020). Shadowbanning was one of the most controversial censorship techniques implemented by platforms, often administered in relation to increasingly sex-averse community guidelines focused on nudity, sexual activity and solicitation (Blunt & Wolf, 2020; Blunt & Stardust, 2021; Paasonen et al., 2019 etc.). As a result, pole dancers were the some of the first shadowbanned groups after sex workers, triggering a widely reported apology that brought further awareness of Instagram's shadowban (Leybold & Nadegger, 2023).

A key characteristic of shadowbanning is that, often, platforms do not notify users it is happening (Are, 2021a). While avoiding notifying those spreading harmful views or content that their posts are being restricted can be useful when running digital spaces, this lack of communication is frustrating and, often, damaging for users, who are made to think their content's lack of views is strictly due its poor quality - a practice dubbed 'blackbox gaslighting' by Cotter (2021). Building on previous work on shadowbanning's impact on sex workers (Blunt et al., 2020), pole dancers (Are, 2021a), influencers, plus size and LGBTQIA+ individuals (Cotter, 2021), this paper provides the first contribution so far evaluating Instagram's most recent attempt at transparency in shadowbanning.

Following repeated backlash against shadowbanning, in December 2022, Instagram announced they were going to notify users their content was being moderated through this technique (Gerken, 2022). By going into the newly created 'Account Status' page in their app's settings, users can now check whether their account is at risk of deletion due to violations of community guidelines, or 'non-recommendable' because, in the platform's words, it may have violated recommendation guidelines (Instagram, n.d.). According to said newly published recommendation guidelines, the Meta-owned app does not recommend: “Content that discusses self-harm, suicide or eating disorders, as well as content that depicts or trivialises themes around death or depression,” content that may depict violence, posts “that may be sexually explicit or suggestive, such as pictures of people in see-through clothing,” or content by accounts the platforms deems non-recommendable, meaning profiles that have repeatedly breached recommendation guidelines (ibid). Following this move, users could also appeal the decision (Gerken, 2022).

Through a qualitative survey amongst 100 pole dancers circulated through the author’s Instagram network and following of almost 30,000, this paper evaluates pole dancers’ relationship with and experience of the newly created Account Status, in order to answer the following research questions:

How has the new Account Status notification of shadowbanning affected pole dancers’ experiences of posting on Instagram?
How effective are Account Status appeals to mitigate or fight the shadowbanning of pole dancing content?

Preliminary findings highlight that while Instagram’s disclosure of their recommendations guidelines has mitigated one problem – the lack of transparency and gaslighting of users (Cotter, 2021), it has also effectively only rebranded a term the company did not use (shadowbanning) into ‘non-recommendable’ (Are, 2021b). Thus, while Instagram may now be disclosing their policies, the effect is the same: swathes of content, and particularly content featuring nudity and LGBTQIA+ expression, is often disproportionately ‘non-recommendable’. Further, users find themselves in a groundhog day of recurring appeals for content that had already been successfully appealed and approved by the platform.

Through a thematic analysis drawing from Daniels’ (1987) idea of invisible labour, or the unpaid and de-valued work women are often expected to perform to enable societal development, this paper highlights various areas of improvement to make the Account Status feature really transparent and useful for pole dancers and nude users alike. Firstly, the inability to reach human moderators to question repeated glitches and appeals makes the experience time-consuming and dehumanising, showing the additional form of work required for women creators and creators at the margins due to the over-moderation of their content, often intersecting with issues of gender, race, sexuality and disability. Secondly, while initial communications with Instagram seemed to hint that the Account Status feature wanted to empower users to take charge of their visibility (Are, 2021b), this empowerment is felt by users rather like a burden, forcing them to appeal into a void while also losing connection and work opportunities due to their constant relegation to the shadows. Lastly, the invisible and often emotional labour performed by pole dancers to avoid shadowbanning ultimately adds information, context and education for other users, cementing their status as platform governance and creative pioneers - a status cemented by the never-ending questioning, conceptualisation and mocking of Instagram’s governance as ‘sexy jail’.

References

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

Are, C. (2021b). Updates on shadowbanning, nudity and ads on Instagram. Blogger On Pole.https://bloggeronpole.com/2021/06/updates-on-shadowbanning-nudity-ads-instagram/

Are, C. (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

Are, C. & Paasonen. S. (2021). Sex in the Shadows of Celebrity. Porn Studies, 8(4) 411-419.

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

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

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. & Stardust, Z. (2021). Automating Whorephobia: Sex, Technology and the Violence of Deplatforming. Porn Studies, 8 (4), 350-366.

Bronstein, C. (2021). Deplatforming sexual speech in the age of FOSTA/ SESTA. Porn Studies, 8(4), 367-380, DOI: 10.1080/23268743.2021.1993972.

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. DOI: 10.1080/1369118X.2021.1994624.

Daniels, A. K. (1987). Invisible work. Social Problems, 34(5), 403–415. https://doi.org/10.2307/800538.

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

Duffy, B. E., & Meisner, C. (2023). Platform governance at the margins: Social media creators’ experiences with algorithmic (in)visibility. Media, Culture & Society, 45(2), 285-304. https://doi.org/10.1177/01634437221111923

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

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

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. Social Media + Society, 8, 3: 20563051221117550. https://doi.org/10.1177/20563051221117552

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.

Instagram (n.d.) Recommendations on Instagram. Help Centre. https://help.instagram.com/313829416281232/.

Leerssen, P. (2023). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. Computer Law & Security Review, 48. https://doi.org/10.1016/j.clsr.2023.105790.

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

PAPER 2: ALGORITHMIC FOLK REMEDIES: HOW CENSORSHIP SHAPED THE ALGORITHMIC IMAGINARY OF QUEER TIKTOK

Alexander Monea
George Mason University

This paper is the final draft of a study I presented in its first iteration at AoIR 2023 (Monea 2023) and thus has significant overlaps with and repurposes material from that presentation. This iteration highlights the ways in which early content moderation strategies continue to shape users' understandings of social media platforms indefinitely into the future. The paper makes this argument through a case study of Queer TikTok's response to early cisheteronormative content moderation policies and practices on the platform, which led to what can be termed 'algorithmic folk remedies' that helped Queer TikTok evade censorship on the platform. This paper argues that these algorithmic folk remedies that resulted from early censorship on the platform continue to shape the way that queer users understand TikTok and create content on the platform. The paper begins with an analysis of theories of algorithmic imaginaries and build connections between them and theories of folk wisdom, folk psychology, and other similar forms of crowdsourced, intuitive, and iterative forms of knowledge production. The paper then moves on to analyze TikTok's content moderation policies and practices, documenting a strong cisheteronormative bias in early moderation efforts on the platform. The paper then moves on to show how these early cisheteronormative content moderation practices shaped users' engagement with the platform as queer users worked to produce algorithmic folk remedies to evade censorship on TikTok. These strategies range from A/B testing to employing 'algospeak' to engaging in cross-platform content strategies, all of which continue to structure content production and everyday use on

Queer TikTok. The paper concludes by extracting some lessons from the case study of Queer TikTok that apply more broadly to global experiences of content moderation, platform governance, algorithmic imaginaries, and algorithmic folk remedies.

TikTok has a troubled history of censoring LGBTQ+ content on its platform. It has censored depictions of homosexuality (e.g., holding hands, touching, kissing), reporting on homosexual groups, content promoting gay rights, and content promoting queerness in general in a number of conservative countries – Turkey being the most famous example (Hern 2019). Research has shown LGBTQ+ related hashtags being suppressed in at least eight languages, including Russian and Arabic (Li 2020; Ryan, Fritz & Impiombato 2020). Many American and Anglophone LGBTQ+ content creators have similarly reported biased moderation of their content, most notably amongst transgender content creators (Akinrinade 2021; Criddle 2020). TikTok has argued that many of these instances were due to their restriction of hashtags associated with 'pornographic searches,' but this simply reifies the pornographication and/or hypersexualization of queer existence so frequently at the center of cisheteronormative censorship. It is worth noting that TikTok's biased content moderation extends to many other socially marginalized groups, including Black content creators (Ghaffary 2021; Rosenblatt 2021), as well as the 'ugly,' poor, and disabled (Biddle, Ribeiro & Dias 2020). As such, TikTok follows in the wake of many other internet platforms in its institution of cisheternormative content moderation policies (Blunt & Stardust 2021; Gehl, Moyer-Horner, & Yeo 2017; Monea 2022).

Within this context, TikTok users are rightfully concerned with the undue censorship of their content and invest their time and energy into avoiding having their content censored, deprioritized, demonetized, or otherwise shadow banned. TikTok has made big promises about opening up its algorithms, policies, and decision-making practices for outside review (Heilweil 2020; Knutson 2020; Matsakis 2020), and it leverages these promises to signal its commitment to 'accountability' and 'transparency' in attempts to mitigate public relations crises – like it coming to light that TikTok censors LGBTQ+ hashtags in eight or more languages (Li 2020). However, TikTok has not actually been very forthcoming with technical details about its algorithms or data about content moderation decisions and everyday users are left with little official or expert knowledge about how TikTok's algorithms are working and what will and will not trigger undue censorship during the content moderation process.

As Abidin notes, "TikTokers have had to rely on repeated attempts, observed patterns, and gut feelings to figure out how the algorithm works, how to please the platform to facilitate their visibility, and how to have their popularity grow" (2020, p. 85). This process is akin to what Bucher calls an 'algorithmic imaginary,' which describes "the way in which people imagine, perceive and experience algorithms and what these imaginations make possible" (2017, p. 31), as well as Bishop's work on 'algorithmic gossip' (2019) and 'algorithmic lore' (2020). Others have described this process as 'algorithmic folklore,' writing:

> **TikTok users have speculated about coded discrimination** on the platform, sharing individual experiences and anecdotal evidence to identify and disrupt the algorithm. Engaging in collective guesswork, these users take to the comments

> section to propose different theories about why the algorithm acts in discriminatory ways. […] By sharing experiences, asking questions, and crowdsourcing answers, **teens are developing an algorithmic folklore** while discerning the potential motivations behind TikTok's software engineering. (Akinrinade 2021)

This is a departure from traditional folk psychology, which explains the ways in which humans perceive, explain, predict, and criticize one another's behavior – largely through the attribution of mental states to others (Hutto & Ravenscroft 2021).
LGBTQ+ users are particularly adept at producing folk knowledge about TikTok's algorithms and content moderation policies, given both the general cisheteronormative bias of online content moderation and TikTok's specific history of censoring LGBTQ+ content on its platform. The practices they employ to identify and avoid cisheteronormative content moderation practices on the platform include (but are not limited to):

- Intentionally using language, keywords, hashtags, and images that they anticipate triggering censorship and cataloging TikTok's responses
- Continually A/B testing the platform's content moderation by posting multiple iterations of the same videos with slight alterations to see which elude the algorithm's content moderation
- Engaging in motivating misspelling and mispronunciation of words understood to trigger censorship (e.g. 'seggs' or 's*ggs' in place of 'sex')
- Tactically covering or obscuring certain parts of the body, background objects, portions of images, etc.
- Leveraging specific audiences, hashtags, and cross platform links to boost LGBTQ+ content
- Utilizing comment spaces and forums on other platforms to collectively produce, collect, archive, and disseminate folk knowledge for LGBTQ+ TikTok users

Queer TikTok is rife with self-censorship and similar attempts to employ algorithmic folk remedies to evade cisheteronormative content moderation on the platform. This is true despite TikTok being popularly understood as the most queer friendly internet platform and actively soliciting queer users. In closing, I argue that this demonstrates two important things about internet platforms: (1) the bar is depressingly low for what counts as 'queer friendly' online, and (2) bad content moderation practices can have a lasting impact on the algorithmic imaginary of a platform's user base.

References

Abidin, C. (2020). Mapping Internet Celebrity on TikTok: Exploring Attention Economies and Visibility Labours. *Cultural Science*, *12*(1): 77-103.

Akinrinade, I. (2021, Jul. 14). Strategic Knowledge: Teens Use "Algorithmic Folklore" to Crack TikTok's Black Box. *Data & Society*. Retrieved from https://points.datasociety.net/strategic-knowledge-6bbddb3f0259 [Accessed Mar. 1, 2023]

Biddle, S., Ribeiro, P.V., Dias, T. (2020, Mar. 16). Invisible Censorship: TikTok Told Moderators to Suppress Posts by “Ugly” People and the Poor to Attract New Users. *The Intercept*. Retrieved from https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/ [Accessed Mar. 1, 2023]

Bishop, S. (2019). Managing Visibility on YouTube through Algorithmic Gossip. *New Media & Society, 21*(11-12): 2589-2606.

Bishop, S. (2020). Algorithmic Experts: Selling Algorithmic Lore on YouTube. *Social Media + Society, 6*(1): 1-11.

Blunt, D., & Stardust, Z. (2021). Automating Whorephobia: Sex, Technology and the Violence of Deplatforming: An Interview with Hacking//Hustling. *Porn Studies*, *8*(4), 350-366.

Bucher, T. (2017). The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms. *Information, Communication & Society, 20*(1): 30-44.

Criddle, C. (2020, Feb. 12). Transgender Users Accuse TikTok of Censorship. *BBC News*. Retrieved from https://www.bbc.com/news/technology-51474114 [Accessed Mar. 1, 2023]

Gehl, R. W., Moyer-Horner, L., & Yeo, S. K. (2017). Training Computers to See Internet Pornography: Gender and Sexual Discrimination in Computer Vision Science. *Television & New Media*, *18*(6), 529-547.

Ghaffary, S. (2021, Jul. 7). How TikTok’s Hate Speech Detection Tool Set Off a Debate About Racial Bias on the App. *Vox*. Retrieved from https://www.vox.com/recode/2021/7/7/22566017/tiktok-black-creators-ziggi-tyler-debate-about-black-lives-matter-racial-bias-social-media [Accessed Mar. 1, 2023]

Heilweil, R. (2020, Jul. 29). Why it Matters that TikTok Wants to Reveal its Algorithms. *Vox*. Retrieved from https://www.vox.com/recode/2020/7/29/21346758/tiktok-for-you-algorithm-transparency-instagram-antitrust [Accessed Mar. 1, 2023]

Hern, A. (2019, Sep. 26). TikTok’s Local Moderation Guidelines Ban Pro-LGBT Content. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2019/sep/26/tiktoks-local-moderation-guidelines-ban-pro-lgbt-content [Accessed Mar. 1, 2023]

Hutto, D. & Ravenscroft, I. (2021). Folk Psychology as a Theory*. Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/folkpsych-theory/ [Accessed Mar. 1, 2023]

Knutson, J. (2020, Mar. 11). TikTok Plans Los Angeles “Transparency Center” to Assuage Critics. *Axios*. Retrieved from https://www.axios.com/2020/03/11/tiktok-los-angeles-china [Accessed Mar. 1, 2023]

Li, J. (2020, Sep. 8). TikTok is Suppressing LGBT Content in Eastern Europe and the Middle East. *Quartz*. Retrieved from https://qz.com/1900530/tiktok-shadow-bans-lgbt-hashtags-in-russian-and-arabic [Accessed Mar. 1, 2023]

Matsakis, L. (2020, Jun. 18). TikTok Finally Explains How the 'For You' Algorithm Works. *WIRED*. Retrieved from https://www.wired.com/story/tiktok-finally-explains-for-you-algorithm-works/ [Accessed Mar. 1, 2023]

Monea, A. (2023, October). Cruising TikTok: Using Algorithmic Folk Knowledge to Evade Cisheteronormative Content Moderation. Paper (or panel) presented at AoIR2023: The 24thAnnual Conference of the Association of Internet Researchers. Philadelphia, PA, USA: AoIR. Retrieved from http://spir.aoir.org

Rosenblatt, K. (2021, Feb. 9). Months after TikTok Apologized to Black Creators, Many Say Little Has Changed. *NBC News*. Retrieved from https://www.nbcnews.com/pop-culture/pop-culture-news/months-after-tiktok-apologized-black-creators-many-say-little-has-n1256726 [Accessed Mar. 1, 2023]

Ryan, F., Fritz, A., & Impiombato, D. (2020). *TikTok and WeChat: Curating and Controlling Global Information Flows*. Policy Brief Report No. 37/2020. Canberra, Australia: Australian Strategic Policy Institute. Retrieved from https://www.aspi.org.au/report/tiktok-wechat [Accessed Mar. 1, 2023]

PAPER 3: MODERATE, MANAGE, MARKET: CONTENT MODERATION AS WORKER MANAGEMENT ON ADULT LABOUR PLATFORMS

Hanne Stegeman
University of Amsterdam

Rébecca Franco
University of Amsterdam

Introduction

Content moderation on adult platforms, like algorithmic management, is a type of worker management. Research on content moderation more generally focuses on the balance between the need to limit harmful content and censorship, and the different private and public actors and their interests involved in these processes (Griffin 2023; Deflem & Silva 2021; McChesney 2013; Gorwa et al 2020). On the flipside, examinations of how platforms manage platform workers tend to focus on employment status classification and algorithmic rankings and ratings (Stark & Pais 2020; Möhlman et al., 2021), and less so on the use of content moderation and user-verification as a management strategy. The separation of these fields of research is reflected in policy debates and regulations. For instance, the UK Online Safety Bill and the EU Digital Services Act deal with content moderation but do not contend with its consequences for content creators as platform workers, while the EU Platform Work Directive regulates platform work and AI management, but not content moderation.

For sex workers and people who create sexual content, however, the (over)moderation of sexual content affects them in ways that closely resemble workers' management. Since platform content moderation disproportionately focusses on sexual content (Gillespie, 2018), some of the effects, experiences and issues with moderation efforts are most prominent in relation to the regulation of this content. Sex workers rely on both adult and mainstream platforms to carry out their work of streaming and selling sexual content. Discussions on the (over)moderation of sexual content have, so far, focussed on representational (Southerton et al., 2020), community-forming (Blunt & Stardust, 2021), income (Hamilton et al., 2022; Ma & Kou, 2021) and emotional (Are & Briggs, 2023) consequences, but less so on how content moderation is experienced as a top-down management technique imposed by platforms. Taking sex work, and particularly erotic webcam streamers, as a case study, this paper contends that content moderation does not just manage "permissable platform content", it also manages platform workers.

Connecting moderation and management

Previous work on content moderation has increasingly focussed on the ways in which sexual content is moderated by platforms. Scholars have observed a general move towards the 'deplatforming' of sex (Blunt et al., 2021; Tiidenberg & Van Der Nagel, 2020), because platforms equate sexual and harmful content in their attempt to avoid legal liabilities, protect their advertisement revenue, and appease relationships with payment processors (Ruberg, 2020; Griffin 2023; Tusikov 2019). Consequences of this have been detailed in various areas. The deplatforming of sex has detrimental consequences, such as the loss of income, workspaces and safety strategies (Blunt et al., 2021; Hamilton et al., 2022). Adult platforms strictly moderate permissible content in ways that leave workers vulnerable to loss of their accounts and revenue (Stegeman, 2021). Online sex workers, such as erotic webcam performers, exemplify what happens when labor platforms are subject to intense moderation.

Because webcam performers are digital workers (Rand, 2019), this case study bridges content moderation and algorithmic gig labour management literature. Algorithms on many gig platforms are responsible for assigning tasks, connecting clients and workers and ultimately influence incomes (Wood et al., 2019). This literature has primarily dealt with how ratings, past jobs and other metrics are extracted by platforms to rank and manage workers (Stark & Pais, 2020; Wood et al., 2019). How content moderation manages workers is underexplored. The experiences of erotic webcam performers, subject to the precarious circumstances of digital work and intense content moderation, showcase the ways in which content moderation is also worker management.

This paper combines perspectives on content moderation and management from multiple angles. Findings are based on: 1. Document analysis of webcamming platforms' terms and conditions, 2. sixteen semi-structured expert interviews with platform insiders, content moderation services, law firms and performers engaged in advocacy, 3. fieldnotes from three adult industry conferences in the US, Romania and the Netherlands, and 4. in-depth interviews with 67 webcam performers from the

Netherlands, Romania and the UK. Together, these methods present a holistic picture of how platforms moderate and manage, and how workers experience and resist this.

Findings

We found that platform interests in controlling workers converge with content moderation requirements, which affects webcam performers. Platforms willfully combine rules and content moderation strategies that are aimed at stopping harmful content with rules that are aimed to serve the business interest of the platform in its relationship with individual performers. For example, platforms' terms and conditions show that platforms instrumentalize the necessity to moderate illegal content to enforce strict management of workers through, for instance, penalty systems for performers that are ostensibly aimed at 'making the site safer'. Moreover, platform representatives argued that they keep some of the content rules deliberately vague in order to allow for implementing such guidelines in ways that serve platform interests while incorporating punitive measures for performers. Some of these forms of control are increasingly backed by AI-powered content moderation systems. The same systems that are required to moderate harmful content also surveil and punish performers who attempt to redirect clients to platforms with higher pay-out rates.

These mechanisms of control through content moderation directly affect webcam performers, who, quite often, outline their own experiences with content moderation as feeling like top down instructions on how to conduct their work. Webcam performers described content rules that have very little to do with content moderation directly influencing their ways of working. One Romanian performer (B) clearly describes this for regulations they experienced on LiveJasmin: "No I stopped working with Jasmin because they had these complicated rules, very strict and they'd modify the percentage of payout you'd get, like, if you don't have the perfect background and the perfect lighting we'll just drop you to 25%". Very clearly, the commercial interests of platforms are also reflected in what a Dutch performer (E) describes as a platform's most important rule "you can't promote other sites, or links, or Skypeshows or whatever, basically anything that would allow you to make money without the platform itself". Here the opacity of how platforms actually enforce their regulations also causes performers to be extra careful and wary: "you have no idea of what content that you're gonna put up is gonna flag" (performer D, UK).

Showing the convergence of content moderation and management strategies in the adult webcamming industry, this paper argues that content moderation creates a de facto managerial relationship between platform and worker. Concerns about platform workers' rights should therefore include an evaluation of the process and effects of content moderation.

References

Are, C., & Briggs, P. (2023). The Emotional and Financial Impact of De-Platforming on Creators at the Margins. Social Media + Society, 9(1), 20563051231155103. https://doi.org/10.1177/20563051231155103

Blunt, D., Duguay, S., Gillespie, T., Love, S., & Smith, C. (2021). Deplatforming Sex: A roundtable conversation. Porn Studies, 8(4), 420–438. https://doi.org/10.1080/23268743.2021.2005907

Blunt, D., & Stardust, Z. (2021). Automating whorephobia: Sex, technology and the violence of deplatforming. Porn Studies, 0(0), 1–17. https://doi.org/10.1080/23268743.2021.1947883

Deflem, M., & Silva, D. M. (Eds.). (2021). Media and law: between free speech and censorship. Emerald Publishing Limited.

Gillespie, T. (2018). Custodians of the Internet : Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven: Yale University Press

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

Griffin, R. (2023). From brand safety to suitability: advertisers in platform governance. Internet policy review, 12(3). DOI: 10.14763/2023.3.1716.

Hamilton, V., Barakat, H., & Redmiles, E. M. (2022). Risk, Resilience and Reward: Impacts of Shifting to Digital Sex Work. arXiv:2203.12728 [Cs]. http://arxiv.org/abs/2203.12728

Ma, R., & Kou, Y. (2021). "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 429:1-429:25. https://doi.org/10.1145/3479573

McChesney R.W. (2013). Digital Disconnect: How Capitalism Is Turning the Internet against Democracy. New York: The New Press.

Möhlmann, Mareike, Lior Zalmanson, Ola Henfridsson, and Robert Wayne Gregory. 2021. "Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control." MIS Quarterly 45(4), 1999–2022. doi:10.25300/MISQ/2021/15333.

Rand, H. M. (2019). Challenging the Invisibility of Sex Work in Digital Labour Politics. Feminist Review, 123(1), 40–55. https://doi.org/10.1177/0141778919879749

Ruberg, B. (2020). "Obscene, pornographic, or otherwise objectionable": Biased definitions of sexual content in video game live streaming. New Media & Society, 1461444820920759. https://doi.org/10.1177/1461444820920759

Southerton, C., Marshall, D., Aggleton, P., Rasmussen, M. L., & Cover, R. (2020). Restricted modes: Social media, content classification and LGBTQ sexual citizenship. New Media & Society, 1461444820904362. https://doi.org/10.1177/1461444820904362

Stark, D., & Pais, I. (2020). Algorithmic Management in the Platform Economy. Sociologica, 14(3), 47–72. https://doi.org/10.6092/issn.1971-8853/12221

Tiidenberg, K., & Van Der Nagel, E. (2020). WHAT'S AT STAKE WHEN SEX IS DEPLATFORMED? AoIR Selected Papers of Internet Research. https://doi.org/10.5210/spir.v2020i0.11348

Wood, A. J., Graham, M., Lehdonvirta, V., & Hjorth, I. (2019). Good Gig, Bad Gig: Autonomy and Algorithmic Control in the Global Gig Economy. Work, Employment and Society, 33(1), 56–75. https://doi.org/10.1177/0950017018785616

PAPER 4: CULTIVATING BRAVE SPACES: ALTERNATIVE APPROACHES TO SEXUAL CONTENT MODERATION

Zahra Stardust
Queensland University of Technology

Digital platforms struggle with the question of how to address harassment, abuse and non-consensual content whilst still facilitating consensual sexual expression. Social media policies on sex are often restrictive (Albury, 2018; Paasonen et al., 2019; Tiidenberg and van der Nagel, 2020), making private, arbitrary, unaccountable decisions about the kinds of sexualities visible online (Stardust, 2018). Platforms have actively shadowbanned, demoted, de-monetised, suspended and deplatformed sex workers and LGBTQ+ folk (Blunt et al., 2020; Monea, 2022; Are, 2022) and social media rules around sex and nudity continue to impact user safety and wellbeing (Southerton et al., 2021). At the same time, a suite of legislation (such as the Fight Online Sex Trafficking Act 2018 in the US, the Online Safety Act 2021 in Australia, and the Online Safety Act 2023 in the UK) now incentivises platforms to automate efforts to remove sexual content, with threat of civil and criminal penalties.

Civil society groups are actively engaging with sexual content moderation. Many have critiqued the discriminatory algorithms, sex-negative policies, extractive business models, surveillance practices and lack of accountability of both firms and governments. Stakeholders have responded to the over-capture of consensual sexual expression, sex education, harm reduction and public health material through a range of measures: manifestos, art projects, written submissions, media campaigns, community research and public interest litigation. Users themselves contest top-down moderation, using creative language to circumvent algorithms that detect sexual solicitation. Among this, generative work is underway among alternative, independent collectives and cooperatives, who are designing new spaces, ethical standards and governance mechanisms.

This empirical study examined how independent small-scale platforms and cooperatives moderate sexual content in ways that move beyond the ethics and practices of corporations and governments. It involved 10 one-hour qualitative interviews with new or emerging platforms that profess to take an alternative approach to sexual content moderation. Mapping approaches between Australia, the United Kingdom and the United States, the study sought to understand how such platforms cultivate consent culture, promote user safety, protect user privacy, support diverse sexualities and value sexual content creators.

Platforms interviewed included Lips Social (a sexual expression site for artists, activists, LGBTQ communities and women, without ‘biased censorship’), Assembly Four (who founded the sex worker-friendly version of Twitter, Switter), Mint Stars (an ‘ethical and inclusive’ subscription site for adult models using Non-Fungible-Tokens), Make Love Not Porn (a streaming site that describes itself as ‘the safest place on the Internet’), and Peep.Me (a sex worker cooperative with the aim of ‘exit to community’). Interviews covered how platforms determine their community standards, decision-making about acceptable sexual content, approaches to tagging and curation, balancing identity verification against user privacy, preventing non-consensual content, human versus automation, and navigating the regulatory environment.

The platforms differentiated their approaches in multiple ways, including through their business models, revenue streams, profit sharing and payouts. They were often built upon the insights of sex workers, featured sex worker leadership, and described being accountable to sex worker communities. They took queer and sex positive approaches to developing community standards, with conceptualisations of harm and safety that differed from legal or risk-averse standards. Some featured bottom-up tagging practices that associated more nuanced meanings with bodies and identities. Some sought to cultivate ‘brave spaces’ for users to unlearn internalised oppressions, and pioneer an educative, transformative justice approach in addressing problematic content. They prioritised valuing sexual content creators, taking lower commissions and donating profits to local sex worker projects.

The platforms studied were often founded in specific legal-techno-political environments that necessitated mutual aid and cooperation: some described how the coalescing of the Black Lives Matter movement with the peak of the COVID-19 pandemic in 2020 was a catalyst for them, especially given that many of their local escort directories had been raided by the FBI and other alternatives like Only Fans were “throwing sex workers under the bus.” Other platforms described how they turned to their offline experiences to think about safety in virtual worlds. For example, Val, the non-binary Latina head of community at Lips Social described how their experience volunteering as a safer space officer in queer night clubs or working as crew on feminist porn sets provided foundational training for thinking about the ethics of governance.

However, the interviews also highlighted how firms and states continue to shape what’s possible in moderating content. While some were deliberately founded on cryptocurrencies, other platforms had shut down citing onerous legislation, threat of prosecution and payment discrimination. Art projects like Only Bans (where users play the role of a sex worker being doxxed and deplatformed) and eViction! (a 12hr, pop-up, self-destructing peepshow) had emerged to draw attention to the economic violence of mainstream sexual content moderation. These platforms still faced difficult challenges – identifying the limits of acceptable sexual content (including where the criminal law had overstepped to prohibit consensual kinks), how to design non-discriminatory algorithms, modelling cooperative decision-making (as they grew in size and scale), and finding options for less data-intensive verification methods (ensuring consensual content, achieving user safety and legal compliance without mass surveillance).

Of the ten platforms interviewed, four had shut down or paused their operations indefinitely, citing the difficulties of bringing their radical visions to life in the current regulatory environment. The platforms – Switter, Sex School, Body of Workers and Peep.Me – expressed their deep sense of grief at losing such important political projects (worker cooperatives! live action porn literacy! sex worker socials!) and their frustration in trying to foreground ethics, accountability and justice in a climate that prioritized speed, scale and surveillance. It's an eternal struggle", said Lina Bembe, a queer feminine migrant of colour and part of the core performer team at Sex School. "It just boils down to – there's no place for platforms to exist."

This study reminds us that there are indeed alternative possibilities for governing sexual content on digital platforms, approaches that improve upon the sexual ethics of both governments and corporations. For such possibilities to flourish, small, local cooperatives must be supported to experiment in imagining different economies of value and relationships to sex, media and community. However, a swathe of regulations (incentivising automation, requiring surveillance, mandating heteronormativity and restricting finance) render it difficult for such projects to even survive, let alone thrive. Law and policy reform is therefore necessary to ensure that sexual content moderation operates not simply as a synonym for 'detection and remove' but instead works to actively cultivate more ethical, just and equitable sexual societies.

References

Albury, K. 2018, 'Sexual Expression in Social Media', p.p. 444-462 in J Burgess, AE Marwick and T Poell (eds), *The Sage Handbook of Social Media*, Sage, London.

Are, C. 2022. 'The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram.' *Feminist Media Studies*. 22:8.

Blunt, D, Wolf A, Coombes E and Mullin S, 2020. 'Posting into the Void: Studying the impact of shadowbanning on sex workers and activists.' *Hacking//Hustling*.

Monea, A. 2022. *The Digital Closet: How the Internet Became Straight*: MIT Press.

Paasonen, S, Jarrett K, Light B, 2019. *#NSFW: Sex, Humor and Risk in Social Media*. MIT Press, Cambridge.

Tiidenberg K and van der Nagel E, 2020. *Sex and Social Media*. Emerald.

Stardust, Z. 2018. 'Safe for Work: Feminist Porn, Corporate Regulation and Community Standards' pp 155-179, in C Dale and R Overell (eds) *Orienting Feminism: Media, Activism and Cultural Representation*, Palgrave Macmillan.

Southerton et al, 2021. ‘Restricted Modes: Social Media, Content Classification and LGBTQ Sexual Citizenship’ 23 *New Media and Society* 920.