Selected Papers of #AoIR2023:
The 24th Annual Conference of the
Association of Internet Researchers
Philadelphia, PA, USA / 18-21 Oct 2023

THE ALGORITHMIC MODERATION OF SEXUAL EXPRESSION: PORNHUB, PAYMENT PROCESSORS AND CSAM

Maggie MacDonald
University of Toronto

Context

Pornography platforms are increasingly required by payment processors to apply algorithmic tools in their content management systems (Gurriell 2021). Particular demands on adult merchants are not proportional to harmful content found on these sites, but a response to the widespread association of pornography with risk (Paasonen et al. 2019; Tiidenberg and van der Nagel 2020). This sex-negative framing is part of a larger trend identified by researchers as the deplatforming of sexual expression (Tiidenberg 2021; Van Dijck et al. 2021; Bronstein 2021; Spišák et al. 2021).

Decades of antipornography campaigning have successfully conflated the legal and regulated porn industry with abuse, nonconsensual content and human trafficking (Webber and Sullivan 2018; Burke and MillerMacPhee 2021; McKee and Lumby 2022). Amplified through uncritical journalism and policy, antiporn claims leave the industry routinely scapegoated as the worst perpetrator of child sexual abuse material (CSAM), as exemplified in 2020 when Pornhub faced allegations of profiting from CSAM circulation. The scandal and subsequent public outrage led to massive service changes on the platform, a parliamentary investigation, and swift demonetization by VISA and Mastercard (Webber et al. 2023).

Harmful content unquestionably circulated on Pornhub, but this problem is not unique to porn platforms. Exponentially more incidents are reported across social media sites which face no financial embargoes over CSAM.[1] Defying the logic of targeting porn to mitigate harm, data instead suggests that financial firms assess merchant risk around public relations interests. Conflation of porn with harm encourages firms to “selectively construct matters of concern” related to risk and safety (Gillett et al. 2022). Reifying

[1] The private nonprofit operating the centralised mandatory reporting system for CSAM, The National Centre for Missing and Exploited Children (NCMEC), received 29,309,106 reports of CSAM possession, manufacture, and distribution in 2021. While 3,393,654 of these reports applied to Instagram and 22,118,952 to Facebook, only 9029 reports of CSAM applied to Pornhub - just 0.02% of Meta’s share.

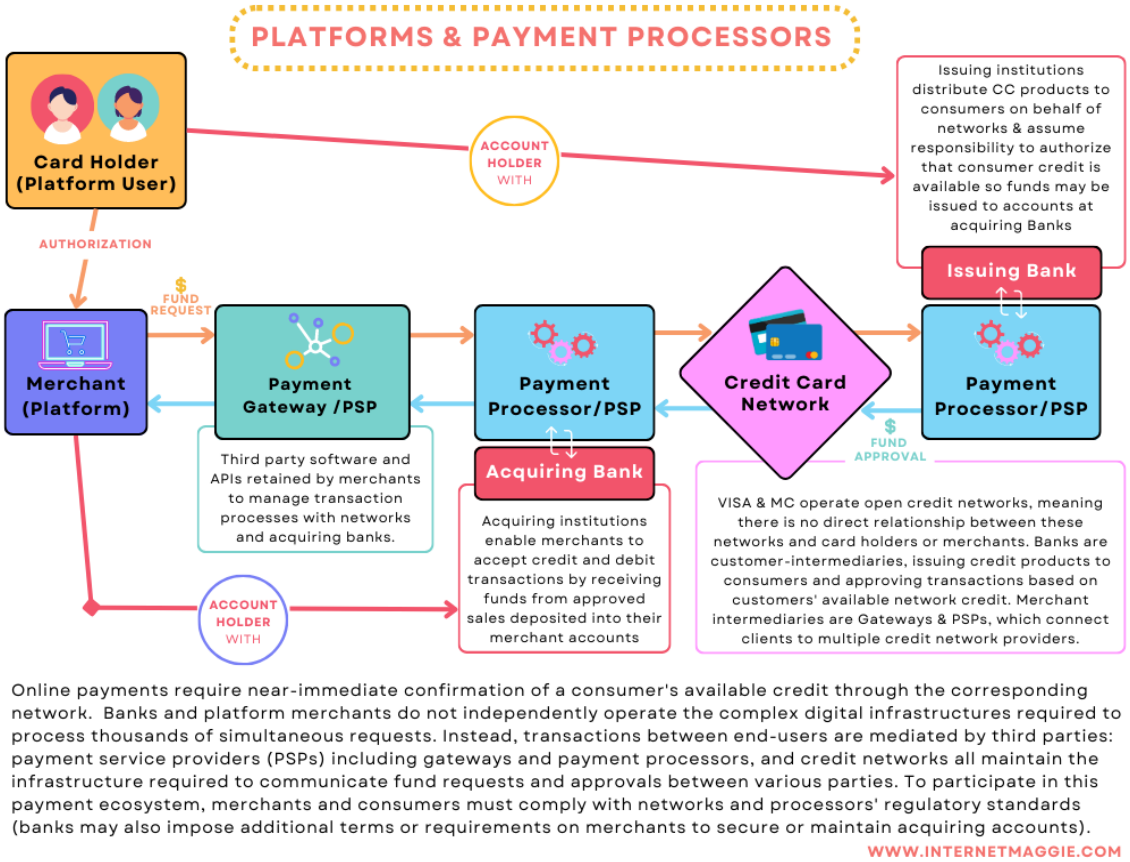
Suggested Citation: MacDonald, Maggie. 2023. *The Algorithmic Moderation of Sexual Expression: Pornhub, Payment Processors and CSAM.* Paper presented at AoIR2023: The 24th Annual Conference of the Association of Internet Researchers. Philadelphia, PA, USA: AoIR. October 20, 2023. Retrieved from http://spir.aoir.org.

whorephobic PR frameworks into technical processes “deputizes private actors to police users whether or not the activity is criminalized”, and perpetuates discrimination by design (Stardust et al. 2023, 137). Responding to Stardust et al.'s call for further empirical study of these systems, I analyse algorithmic moderation tools on Pornhub to ask: what standards are defined by financial firms, how are these enforced, and what effects does this arrangement have on pornographic content?

Methods & Analysis

I take a three-pronged approach. First, mapping financial infrastructures clarifies the enforcement of rules by intermediary business partners.

Along with formal global, state, regional and self-regulation, platforms are subject to significantly less-clear co-regulation by powerful stakeholders (Gorwa 2019). Operators, users, partner firms, third-party services, policymakers, advocates and more all have unequal influence in negotiating and setting standards (Nieborg et al. 2021). As the primary engine of global digital commerce, credit networks - like platforms - have power to establish dependencies and enforce standards among partners.



[Figure 1: business partnerships of digital credit infrastructures]

Transaction ecosystems are complicated. Credit firms do not interact with platforms directly, but through intermediary banks and payment service providers (PSPs). PSPs include both processors—which convey transactions between credit networks and issuing or acquiring banks (the largest include JPMorgan Chase, CitiBank, Wells Fargo

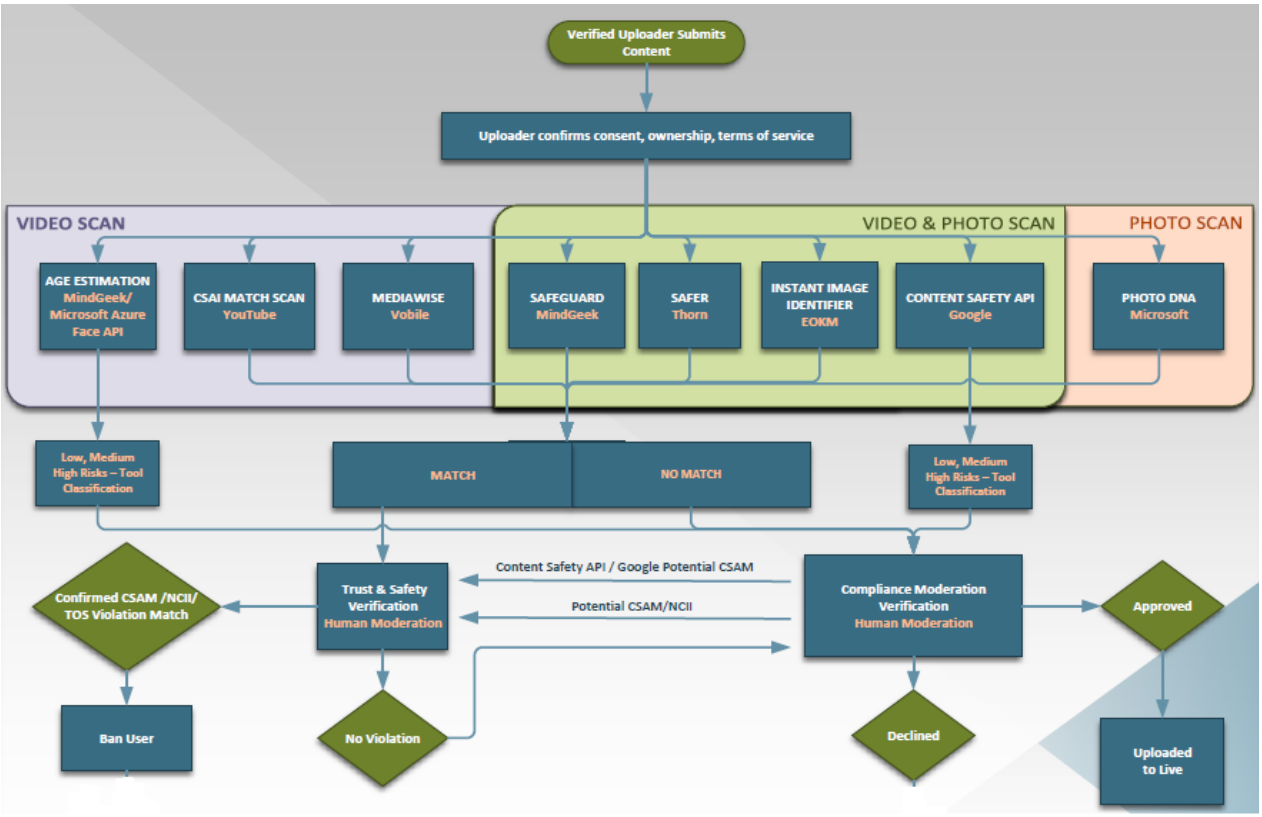
and Fiserv)—and gateways which operate as point-of-sale, encrypting and verifying customer information before sending requests to processors (Shopify, Helcim, MerchantOne, Apple Pay, and Amazon Pay). Many operate as both (PayPal, Square, Stripe, Venmo). PSPs and banks must enforce credit network standards or risk fines and imperil their own valuable partnership contracts. VISA and Mastercard are two of the largest credit providers, holding a duopoly over financial infrastructures essential to platform success. While not formally recognized as regulators, these companies are imbued with powers of co-governance (Gorwa et al. 2020). Identified as essential infrastructures for financial inclusion, this opaque system is a chokepoint allowing denial of services with limited culpability (O'Brien and Reitman 2020).

Next, a close reading of merchant agreements identifies clauses calling for algorithmic intervention. Coded as a reputational liability, pornography becomes a symbolic, rather than substantial, source of harm. Adult merchants are subsequently flagged high-risk and subject to enhanced compliance measures (Free Speech Coalition 2023). For example, Mastercard specifies “automated tools and solutions are not only permissible but recommended”, requiring high-risk merchants “review all content before it is published, and have systems in place for real-time monitoring of livestreams” (2021). VISA insists merchants “must safeguard against risks that may negatively affect their brand or reputation” and compliance recommendations include “machine learning to review data” (2021). Notably absent are standardised guidelines for acceptable systems, providers, or centralised bodies managing the burden of compliance. These ill-defined standards shield financial firms from accountability over discrimination, but incentivize partner firms to refuse ‘risky’ clients (Tusikov 2021).

Finally, a taxonomy of Pornhub tools examines a dominant porn platform’s response to these standards.

Classifier - Hasher	PhotoDNA (Microsoft) Instant II (EOKM) Safer (Thorn) MediaWise (Vobile)
Classifier - Crawler	Project Arachnid (C3P)
Predictor	SafeGuard (Aylo)
CNN	CSAI Match (YouTube) ContentSafety API (Google)

[Figure 2: Pornhub’s moderation tool classifications]



[Figure 3: content moderation process, Pornhub 2022 Transparency Report]

Pornhub applies seven external tools and one proprietary system (SafeGuard) to user-generated content. Classifiers identify known images using a process of metadata similarity detection called hashing. Web-crawlers collect the totality of a website's content, identifying suspicious patterns through aggregated data from keywords and hyperlinks. Predictors mostly rely on visual detection, isolating image features to classify and sort content. Finally, recent developments in Convolutional Neural Networks (CNNs) have hybridised visual and metadata analysis, applying hashing and predictive systems in tandem.

Issue Discussion & Conclusion

Porn platforms are criticised as under-regulated, lacking "algorithmic accountability" in systems that prioritise engagement above all else, but function at a scale making algorithmic intervention necessary (Hunt and McKelvey 2019; Gillespie 2020). With eight tools in place, Pornhub exceeds credit card compliance guidelines, yet remains blacklisted over reputational risk. Positioned as both problem and solution, algorithmic moderation is merely symbolic of increased safety, security and responsibility, but these systems "remain opaque, unaccountable and poorly understood" (Gorwa et al. 2020, page 1).

Forensic, biometric and computer science research supports developer claims around high rates of accuracy (Sanchez et al. 2019; Pereira et al. 2020). However, researchers from platform, pornography and legal, critical race and feminist technoscience studies

express substantial concern around bias in decontextualized application of these systems (Gehl et al. 2017; Buolamwini and Gebru 2018; Gerrard and Thornham 2020; Are 2020; Krishna 2021; Blunt and Stardust 2021; Coombes et al. 2022).

Distinguishing CSAM from legal porn consistently returns false positive rates around 13% in error margins developers dismiss as “not ideal” (Lee et al. 2020, 7). Visual analysis cannot determine consent, lacks accuracy determining age and essentializes complex identity expressions for race and gender (Lee et al. 2020; Scheuerman et al. 2021). Ethical concerns with CSAM training data abound, and generalised datasets reduce accuracy for marginalised subjects, meaning hegemonic white, cis and able bodied content is more accurately identified (Laranjeira et al. 2022). Non-white children are less proximate to rescue from CSAM, and non-white content more prone to errors, thus establishing ‘digital redlining’ in detection systems (Thakor 2018; Tusikov 2021).

Human moderators remain essential, but workers lack appropriate training or resources to manage traumatising content (Mount et al. 2021). Best practices in CSAM reduction require robust structural support focused on prevention, education and legal interventions, generally not offered by employers seeking cost-cutting efficiency through the promise of automation (Kloess et al. 2019; 2021).

Payment processor demands on porn platforms substantiate a case where “scientific knowledge, technological innovation, and corporate profit reinforce each other in deeply entrenched patterns that bear the unmistakable stamp of political and economic power” (Winner 1980, 126). CSAM moderation will not improve through undemocratic co-governance or invocations of safe and neutral algorithms. Applied uncritically, these tools reinforce old prejudices in service of political regimes devaluing sexual expression.

References

Are, Carolina. 2020. “How Instagram’s Algorithm Is Censoring Women and Vulnerable Users but Helping Online Abusers.” *Feminist Media Studies* 20 (5): 741–44. https://doi.org/10.1080/14680777.2020.1783805.

Blunt, Danielle, Stefanie Duguay, Tarleton Gillespie, Sinnamon Love, and Clarissa Smith. 2021. “Deplatforming Sex: A Roundtable Conversation.” *Porn Studies* 8 (4): 420–38. https://doi.org/10.1080/23268743.2021.2005907.

Blunt, Danielle, and Zahra Stardust. 2021. “Automating Whorephobia: Sex, Technology and the Violence of Deplatforming.” *Porn Studies* 8 (4): 350–66. https://doi.org/10.1080/23268743.2021.1947883.

Bronstein, Carolyn. 2021. “Deplatforming Sexual Speech in the Age of FOSTA/SESTA.” *Porn Studies* 8 (4): 367–80. https://doi.org/10.1080/23268743.2021.1993972.

Buolamwini, Joy, and Timnit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In *Proceedings of the 1st Conference*

on Fairness, Accountability and Transparency, edited by Sorelle A. Friedler and Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html.

Burke, Kelsy, and Alice MillerMacPhee. 2021. "Constructing Pornography Addiction's Harms in Science, News Media, and Politics." *Social Forces; a Scientific Medium of Social Study and Interpretation* 99 (3): 1334–62. https://doi.org/10.1093/sf/soaa035.

Coombes, Emily, Ariel Wolf, Danielle Blunt, and Kassandra Sparks. 2022. "Disabled Sex Workers' Fight for Digital Rights, Platform Accessibility, and Design Justice." *Disability Studies Quarterly: DSQ* 42 (2). https://doi.org/10.18061/dsq.v42i2.9097.

Free Speech Coalition and Sex Work CEO. 2023. "Financial Discrimination and the Adult Industry." Free Speech Coalition. https://www.freespeechcoalition.com/banks.

Gehl, Robert W., Lucas Moyer-Horner, and Sara K. Yeo. 2017. "Training Computers to See Internet Pornography: Gender and Sexual Discrimination in Computer Vision Science." *Television & New Media* 18 (6): 529–47. https://doi.org/10.1177/1527476416680453.

Gerrard, Ysabel, and Helen Thornham. 2020. "Content Moderation: Social Media's Sexist Assemblages." *New Media & Society* 22 (7): 1266–86. https://doi.org/10.1177/1461444820912540.

Gillespie, Tarleton. 2020. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7 (2): 2053951720943234. https://doi.org/10.1177/2053951720943234.

Gillett, Rosalie, Zahra Stardust, and Jean Burgess. 2022. "Safety for Whom? Investigating How Platforms Frame and Perform Safety and Harm Interventions." *Social Media + Society* 8 (4): 20563051221144315. https://doi.org/10.1177/20563051221144315.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1): 2053951719897945. https://doi.org/10.1177/2053951719897945.

Gorwa, Robert. 2019. "The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content." *Internet Policy Review* 8 (2). https://doi.org/10.14763/2019.2.1407.

Gurriell, Maryjane. 2021. “Born into Porn but Rescued by Thorn: The Demand for Tech Companies to Scan and Search for Child Sexual Abuse Images.” *Family Court Review* 59 (4): 840–54. https://doi.org/10.1111/fcre.12613.

Hunt, Robert, and Fenwick McKelvey. 2019. “Algorithmic Regulation in Media and Cultural Policy: A Framework to Evaluate Barriers to Accountability.” *Journal of Government Information: An International Review of Policy, Issues and Resources* 9: 307–35. https://doi.org/10.5325/jinfopoli.9.2019.0307.

Kloess, Juliane A., Jessica Woodhams, Helen Whittle, Tim Grant, and Catherine E. Hamilton-Giachritsis. 2019. “The Challenges of Identifying and Classifying Child Sexual Abuse Material.” *Sexual Abuse: A Journal of Research and Treatment* 31 (2): 173–96. https://doi.org/10.1177/1079063217724768.

Kloess, Juliane A., Jessica Woodhams, and Catherine E. Hamilton-Giachritsis. 2021. “The Challenges of Identifying and Classifying Child Sexual Exploitation Material: Moving towards a More Ecologically Valid Pilot Study with Digital Forensics Analysts.” *Child Abuse & Neglect* 118 (August): 105166. https://doi.org/10.1016/j.chiabu.2021.105166.

Krishna, Anirudh. 2021. “Internet.gov: Tech Companies as Government Agents and the Future of the Fight Against Child Sexual Abuse.” *California Law Review* 109 (4). https://lawcat.berkeley.edu/record/1211448.

Laranjeira, Camila, João Macedo, Sandra Avila, and Jefersson A. dos Santos. 2022. “Seeing without Looking: Analysis Pipeline for Child Sexual Abuse Datasets.” *arXiv [cs.CV]*. https://doi.org/10.48550/arXiv.2204.14110.

Lee, Hee-Eun, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. 2020. “Detecting Child Sexual Abuse Material: A Comprehensive Survey.” *Forensic Science International: Digital Investigation* 34 (September): 301022. https://doi.org/10.1016/j.fsidi.2020.301022.

Mastercard. 2021. “Business Risk Assessment and Mitigation (BRAM) Compliance Program - Revised Standards for Specialty Merchant Registration & Requirements for Adult Content Merchants.” AN 5196. Mastercard.

McKee, Alan, and Catharine Lumby. 2022. “Pornhub, Child Sexual Abuse Materials and Anti-Pornography Campaigning.” Porn Studies, June, 1–13. https://doi.org/10.1080/23268743.2022.2083662.

Mount, David, Lorraine Mazerolle, Renee Zahnow, and Leisa James. 2021. “Triaging Online Child Abuse Material: Testing a Decision Support Tool to Enhance Law Enforcement and Investigative Prioritisation.” *Policing: An International Journal* 44 (4): 628–42. https://doi.org/10.1108/PIJPSM-02-2021-0020.

Nieborg, David B., Thomas Poell, and Brooke Erin Duffy. 2021. “Analyzing Platform Power in the Culture Industries.” *Spirales*, September. https://doi.org/10.5210/spir.v2021i0.12219.

O’Brien, Danny, and Rainey Reitman. 2020. “Financial Censorship: Visa and Mastercard Are Trying to Dictate What You Can Watch on PornHub.” Electronic Frontier Foundation. December 14, 2020. https://www.eff.org/issues/financial-censorship.

Paasonen, Susanna, Kylie Jarrett, and Ben Light. 2019. *NSFW: Sex, Humor, and Risk in Social Media*. MIT Press.

Pereira, Mayana, Rahul Dodhia, Hyrum Anderson, and Richard Brown. 2020. “Metadata-Based Detection of Child Sexual Abuse Material.” *arXiv [cs.LG]*. https://doi.org/10.48550/arXiv.2010.02387.

Pornhub. 2022 “Transparency Report.” Accessed October 7, 2023. https://help.pornhub.com/hc/en-us/articles/14666334117267-2022-Transparency-Report-.

Sanchez, Laura, Cinthya Grajeda, Ibrahim Baggili, and Cory Hall. 2019. “A Practitioner Survey Exploring the Value of Forensic Tools, AI, Filtering, & Safer Presentation for Investigating Child Sexual Abuse Material (CSAM).” *Digital Investigation* 29 (July): S124–42. https://doi.org/10.1016/j.diin.2019.04.005.

Scheuerman, Morgan Klaus, Alex Hanna, and Emily Denton. 2021. “Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development.” *Proceedings of. ACM Hum.-Comput. Interact.*, 317, 5 (CSCW2): 1–37. https://doi.org/10.1145/3476058.

Spišák, Sanna, Elina Pirjatanniemi, Tommi Paalanen, Susanna Paasonen, and Maria Vihlman. 2021. “Social Networking Sites’ Gag Order: Commercial Content Moderation’s Adverse Implications for Fundamental Sexual Rights and Wellbeing.” *Social Media + Society*. https://doi.org/10.1177/20563051211024962.

Stardust, Zahra, Danielle Blunt, Gabriella Garcia, Lorelei Lee, Kate D’Adamo, and Rachel Kuo. 2023. “High Risk Hustling: Payment Processors Sexual Proxies and Discrimination by Design.” *City University of New York Law Review* 26 (1): 57–138.

Thakor, Mitali. 2018. “Digital Apprehensions: Policing, Child Pornography, and the Algorithmic Management of Innocence.” *Catalyst* 4 (1): 1–16. https://doi.org/10.28968/cftt.v4i1.29639.

Tiidenberg, Katrin, and Emily van der Nagel. 2020. Sex and Social Media. Emerald Publishing Limited.

Tiidenberg, Katrin. 2021. “Sex, Power and Platform Governance.” *Porn Studies* 8 (4): 381–93. https://doi.org/10.1080/23268743.2021.1974312.

Tusikov, Natasha. 2021. “Censoring Sex: Payment Platforms’ Regulation of Sexual Expression.” In *Media and Law: Between Free Speech and Censorship*, 63–79. Sociology of Crime, Law and Deviance. Emerald Publishing Limited. https://doi.org/10.1108/s1521-613620210000026005.

Van Dijck, José, Tim de Winkel, and Mirko Tobias Schäfer. 2021. “Deplatformization and the Governance of the Platform Ecosystem.” *New Media & Society* 1 (17): 14614448211045662. https://doi.org/10.1177/14614448211045662.

VISA. 2021. “Global Brand Protection Program (GBPP) Guide for Acquirers- Payment Facilitator and Marketplace Risk Guide.” VISA.

Webber, Valerie, and Rebecca Sullivan. 2018. “Constructing a Crisis: Porn Panics and Public Health.” *Porn Studies* 5 (2): 192–96. https://doi.org/10.1080/23268743.2018.1434110.

Webber, Valerie, Maggie MacDonald, Stefanie Duguay, and Fenwick McKelvey. 2023. “Pornhub and Policy: Examining the Erasure of Pornography Workers in Canadian Platform Governance.” *Canadian Journal of Communication* 48 (2): 381–404. https://doi.org/10.3138/cjc.2022-0044.