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.

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.

<table>
<thead>
<tr>
<th>Classifier - Hasher</th>
<th>PhotoDNA (Microsoft)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Instant II (EOKM)</td>
</tr>
<tr>
<td></td>
<td>Safer (Thorn)</td>
</tr>
<tr>
<td></td>
<td>MediaWise (Vobile)</td>
</tr>
<tr>
<td>Classifier - Crawler</td>
<td>Project Arachnid (C3P)</td>
</tr>
<tr>
<td>Predictor</td>
<td>SafeGuard (Aylo)</td>
</tr>
<tr>
<td>CNN</td>
<td>CSAI Match (YouTube)</td>
</tr>
<tr>
<td></td>
<td>ContentSafety API (Google)</td>
</tr>
</tbody>
</table>

[Figure 2: Pornhub’s moderation tool classifications]
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


