

Chapter 43

Governing Image-Based Sexual Abuse: Digital Platform Policies, Tools, and Practices

Nicola Henry and Alice Witt


Abstract

The nonconsensual taking or sharing of nude or sexual images, also known as “image-based sexual abuse,” is a major social and legal problem in the digital age. In this chapter, we examine the problem of image-based sexual abuse in the context of digital platform governance. Specifically, we focus on two key governance issues: first, the governance of platforms, including the regulatory frameworks that apply to technology companies; and second, the governance by platforms, focusing on their policies, tools, and practices for responding to image-based sexual abuse. After analyzing the policies and practices of a range of digital platforms, we identify four overarching shortcomings: (1) inconsistent, reductionist, and ambiguous language; (2) a stark gap between the policy and practice of content regulation, including transparency deficits; (3) imperfect technology for detecting abuse; and (4) the responsabilization of users to report and prevent abuse. Drawing on a model of corporate social responsibility (CSR), we argue that until platforms better address these problems, they risk failing victim-survivors of image-based sexual abuse and are implicated in the perpetration of such abuse. We conclude by calling for reasonable and proportionate state-based regulation that can help to better align governance by platforms with CSR-initiatives.

Keywords: Digital platforms; platform governance; image-based sexual abuse; nonconsensual pornography; content moderation; corporate social responsibility

The Emerald International Handbook of Technology-Facilitated Violence and Abuse, 749–768

Copyright © 2021 Nicola Henry and Alice Witt

 Published by Emerald Publishing Limited. This chapter is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of these chapters (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at <http://creativecommons.org/licenses/by/4.0/legalcode>.

doi:10.1108/978-1-83982-848-520211054

Introduction

The nonconsensual taking or sharing of nude or sexual images, also known as “image-based sexual abuse” (Henry et al., 2020; McGlynn & Rackley, 2017) or “nonconsensual pornography” (Citron & Franks, 2014; Ruvalcaba & Eaton, 2020), is a major social and legal problem in the digital age. With the development of social media and other networked technologies, which enable over three billion users to generate and instantaneously share content on the internet (Kemp, 2020), image-based sexual abuse is not only rapidly increasing, but also having significant impacts (Henry et al., 2020).

While criminal offenses are an important means to punish perpetrators and provide justice to victim-survivors, criminalization has done little to prevent the scourge of image-based sexual abuse or minimize the harm once images (photographs or videos) are posted online. For example, images can be copied and republished on multiple platforms and devices – in some cases making it virtually impossible to prevent the further spread of images online. Perpetrators are often difficult to identify because of anonymity measures, such as encryption, virtual private networks, and proxy servers that obscure the nature of content, locations of internet traffic, and other information about users and their devices. Moreover, policing for image-based sexual abuse (and cybercrime more generally) is typically resource intensive given that law enforcement agencies often have to work across jurisdictional borders.

In response to the complex challenges raised by harmful online content, governments around the world are introducing new regulatory regimes to attempt to better hold technology companies accountable for hosting harmful content on their platforms. At the same time, technology companies are themselves taking more proactive steps to tackle this problem. In this chapter, we examine the problem of image-based sexual abuse in light of these two forms of governance. In the first section, we focus on the *governance of* digital platforms, examining the introduction of broader governmental and intergovernmental regulatory regimes in a changing landscape, which some have described as a global “techlash” against the major digital platforms (Flew, Martin, & Suzor, 2019, p. 33). In the second section, we examine the *governance by* digital platforms, focusing specifically on the policies, tools, and practices that are being implemented by digital platforms to respond to and prevent image-based sexual abuse.

In the third section, we draw on a model of corporate social responsibility (CSR) to propose ways forward. CSR provides a useful, albeit contested, language to examine the policy and practice of online content moderation or regulation. Although there are different conceptions of CSR, we define it as corporations’ social, economic, legal, moral, and ethical responsibilities to address the harmful effects of their activities. Our conception of CSR is embedded within a social justice framework that locates the rationale for action not solely as a profit- or reputation-building exercise, but one that is also contingent on community values and the “common good.”

We argue that while many digital platforms are taking proactive steps to detect and address image-based sexual abuse, four main shortcomings are evident in their policy approaches. First, some platforms adopt inconsistent, reductionist,

and ambiguous language to describe image-based sexual abuse. Second, although a number of platforms now have an explicit policy position on image-based sexual abuse, there is often a stark gap between the policy and practice of content regulation, as well as a lack of transparency about how decisions are made and what the outcomes of those decisions are. Third, while platforms are increasingly turning to high-tech solutions to either detect or prevent image-based sexual abuse, these are imperfect measures that can be circumvented. And fourth, the onus is predominantly placed on users to find and report image-based sexual abuse to the platforms, which can be retraumatizing and highly stressful.

We contend that because of their governing power, public character, and control of information, digital platforms have an ethical responsibility to detect, address, and prevent image-based sexual abuse on their networks. This is despite the degree of legal immunity that platforms have against harmful content posted by their users under section 230(c) of the United States (US) Communications Decency Act of 1996 (CDA 230). We argue that when platforms govern without sufficient regulatory safeguards in place, such as appeal processes and reasoning practices (Suzor, 2019), they risk failing victim-survivors of image-based sexual abuse and are implicated in the perpetration of image-based sexual abuse.

Governance of Digital Platforms

Also known as “internet intermediaries,” or “online service providers,” digital platforms are nonstate, corporate organizations or entities that facilitate transactions, information exchange, or communications between third parties on the internet (see, e.g., Taddeo & Floridi, 2016). According to Gillespie (2018), digital platforms are “sites and services that host public expression, store it on and serve it up from the cloud, organize access to it through search and recommendation, or install it onto mobile devices” (p. 254). Gillespie (2018) explains that what digital platforms share in common is the hosting and organization of “user content for public circulation, without having produced or commissioned it” (p. 254). While digital platforms might appear to be neutral conduits or proxies for the exchange of online content between third parties, they are never neutral, and have been described as the “new governors” or “superpowers” of the digital age (Klonick, 2018; Lee, 2018). Some commentators argue that technology companies are engaged in illicit forms of digital surveillance, plundering the behavioral data of users to sell to business customers (including political advertisers) for economic profit (e.g., Zuboff, 2019), as well as creating the norms and means through which individual users can engage in “performative surveillance” in the form of tracking, monitoring, and observing other users online (e.g., Westlake, 2008).

In addition to potentially illicit forms of surveillance and data harvesting, one of the key ways platforms govern their networks is by moderating user-generated content. As a form of regulation, content moderation encompasses an array of processes through which platform executives and their employees set, maintain, and enforce the bounds of “appropriate” user behaviors (Witt, Suzor, & Huggins, 2019). The norm is for content moderation to be *ex post*, meaning it is undertaken after a user has posted content, and reactive in response to user flags or reports (Klonick, 2018; Roberts, 2019). This means that platforms generally do not

proactively screen content, decisions about which are thus predominantly made *after* the material is posted. On some platforms, however, automated systems are increasingly playing a more central role in the detection and removal of harmful online content before anyone has the chance to view or share the material (see further discussion below).

There are significant transparency deficits around the ways that different types of content are moderated in practice (Witt et al., 2019, p. 558). It is often unclear, for instance, what material is signaled for removal, how much content is actually removed, and by what means. It is also impossible to determine precisely who removes content (e.g., a platform content moderator or a user) without access to a platform's internal workings (Witt et al., 2019, p. 572). The secrecy around the inner workings of content moderation is reinforced by the operation of contract law, which governs the platform–user relationship, and powerful legal protections under US law (where many platforms are primarily based). Specifically, CDA 230 protects platforms against liability for content posted by third parties. Consequently, platforms that host or republish content are generally not legally liable for what their users say or do except for illegal content or content that infringes intellectual property regimes. Indeed, technology companies not only exercise “unprecedented power” over “what [users] can see or share” (Suzor, 2019, p. 8), but also have “broad discretion to create and enforce their rules in almost any way they see fit” (Suzor, 2019, p. 106). This means that decisions around content can be based on a range of factors, including public-facing policies like terms of service, community guidelines, prescriptive guidelines that moderators follow behind closed doors, legal obligations, market forces, and cultural norms of use.

Digital platforms are not, however, completely “lawless” (Suzor, 2019, p. 107). Platforms are subject to a range of laws in jurisdictions around the globe, some of which have the potential to threaten the ongoing stability of the CDA 230 safe harbor provisions. Europe has been described as the “world’s leading tech watchdog” (Satariano, 2018) especially with European regulators taking an “increasingly activist stance toward... digital platform companies” (Flew et al., 2019, p. 34). The European Union’s General Data Protection Regulation (GDPR) and Germany’s NetzDG laws, for instance, can result in significant administrative fines for data protection or security infringements (among other punitive consequences for noncompliance) (see Echikson & Knodt, 2018; *The European Parliament and the Council of the European Union, 2016/679*). There are also many examples of European courts ordering service providers to restrict the types of content users see and how and when they see it (e.g., copyright or defamation lawsuits) (Suzor, 2019, p. 49).

These state-based “regulatory pushbacks” are part of a global “techlash” against the governing powers of digital platforms in recent years (Flew et al., 2019, pp. 33 and 34). At the time of writing this chapter, the United Kingdom had proposed a range of measures in its White Paper on Online Harms, which includes a statutory duty of care that will legally require platforms to stop and prevent harmful material appearing on their networks (Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, 2019). In 2019, Canada released the Digital Charter in Action, which includes 10 key principles designed to ensure the ethical collection, use, and disclosure of data (Innovation, Science and Economic Development Canada, 2019).

Going a step further, after the Christchurch mosque shootings in New Zealand on March 15, 2019, the Australian Federal Government passed the *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth) which gives the Australian eSafety Commissioner powers to issue take-down notices to digital platforms that host abhorrent violent material (AVM). If a service provider fails to remove AVM, they can be subject to prosecution under Australian federal criminal law, among other potential courses of action. Moreover, in 2018, the Australian federal government introduced an innovative civil penalty scheme which prohibits the nonconsensual sharing of intimate images, as well as threatening to share intimate images. Under this scheme, the eSafety Commissioner can issue substantial fines, formal warnings, infringement notices, or take-down notices to individuals and corporations requiring the removal of images within 48 hours.

These domestic and international developments recognize that the decision-making processes of ostensibly “private” digital platforms can have significant impacts on individual users and far-reaching implications for politics, culture, and society (the “public sphere”) more broadly. They also suggest that platform immunity from legal liability for both privacy violations and the hosting of harmful content is diminishing – at least in some jurisdictional contexts.

Digital platforms might then not be completely lawless, but do in practice govern, to use Suzor’s (2019) term, “in a lawless way” (p. 107). Platforms exercise extraordinary power with limited safeguards for users, such as fairness, equality, and certainty, which many Western citizens have come to expect from governing actors (Witt et al., 2019). The result is often a significant gap between platform policies and their governance in practice, as well as a lack of transparency around digital platforms’ decision-making processes.

Governance by Digital Platforms

In this section, we explore an array of policies, tools, and practices that are designed to detect, prevent, and respond to image-based sexual abuse on some of the largest digital platforms. Given the rapid pace of innovation in the technology sector, we selected platforms according to their traffic, market dominance, and their capacity to host image-based sexual abuse content. The sites we selected were predominantly the most popular sites as ranked by the analytics company Alexa (*Alexa Internet, n.d.*). The social media and search engine platforms we examined included Google, YouTube, Facebook, Yahoo!, Reddit, Instagram, Microsoft, Twitter, Flickr, Snapchat, TikTok, and Tumblr. The pornography sites we examined included Pornhub, XVideos, and xHamster. After creating a list of sites, we used the Google search engine to identify each company’s policy documents, including their terms of service, community guidelines, reports, and official blogs. Each document was analyzed to identify specific image-based sexual abuse policies, general policies that could be applicable to image-based sexual abuse, and tools for either detecting, reporting, or blocking content, if any. We also searched for any relevant news articles or blogs on platforms’ responses to image-based sexual abuse content.

Our approach has four main limitations. The first limitation is that we were only able to examine publicly available policy documents. As such, we were not able to examine the undisclosed guidelines that moderators follow behind closed doors or information about the privatized automated systems that digital platforms might use. Second, we carried out our analysis over a three-month period between January and March 2020 and thus we cannot account for any changes in policies, tools, or practices after this time. Third, we did not examine non-English technology companies, nor did we examine the fringe, “rogue,” or underground platforms (e.g., on the Clear Net or Dark Net) where image-based sexual abuse content is being shared and traded (see [Henry & Flynn, 2019](#)).

Finally, we did not seek to empirically investigate the experiences or perspectives of either victim-survivors or platform representatives in relation to content removal or platform policies, tools, and practices. Currently there is a pervasive lack of transparency around platform governance and more research is needed to address this gap. The analysis below, however, provides insight into how select platforms are attempting to address and prevent image-based sexual abuse. Here we focus on three key areas of content moderation: platform policies; reporting options and practices; and technological tools.

Platform Policies on Image-Based Sexual Abuse

The term “revenge porn” came into popular usage in 2011 after widespread media attention to the nonconsensual sharing of nude or sexual images of musicians and sportspersons on the website [IsAnyoneUp.com](#) and the subsequent criminal trial of its founder Hunter Moore ([Martens, 2011](#)). The term, however, is a misnomer because not all perpetrators are motivated by revenge when they share nude or sexual images without consent. Instead they may be motivated for other reasons, such as sexual gratification, monetary gain, social status building, or a desire for power and control ([Citron & Franks, 2014](#); [Henry et al., 2020](#)). The term “revenge porn” has been widely criticized as having victim-blaming, harm-minimizing, or salacious connotations. Scholars, activists, victim-survivors, and practitioners also argue that it fails to capture the complexity and diversity of behaviors involving the use and abuse of nonconsensual nude or sexual images by known and unknown persons alike, using diverse means and methods ([Henry et al., 2020](#); [McGlynn & Rackley, 2017](#); [Powell, Henry, & Flynn, 2018](#)).

Although a small number of digital platforms continue to refer to “revenge porn” in their terms of service or community guidelines, others have adopted alternative terms, such as “nonconsensual pornography,” “involuntary pornography,” or “the nonconsensual sharing of intimate images.” Tumblr’s community guidelines, for instance, state: “Absolutely do not post nonconsensual pornography – that is, private photos or videos taken or posted without the subject’s consent” ([Tumblr, 2020](#), Privacy violations, para 1). Other platforms outline prohibitions against broader forms of online content. For instance, Pornhub’s terms of service explicitly prohibit, among other behaviors, the impersonation of another person, the posting of copyrighted material, content that depicts a person under the age of 18, and content that

is “obscene, illegal, unlawful, defamatory, libellous, harassing, hateful, racially, or ethnically offensive” (Pornhub, 2020, Monitoring and enforcement, para 4). Notably, however, Pornhub does not specify explicit prohibitions against image-based sexual abuse. In their policies, xHamster and XVideos do not specifically mention image-based sexual abuse but instead refer to privacy, abuse, harassment, inappropriate, or illegal content (xHamster, 2020; XVideos, n.d.). TikTok’s Community Policy similarly does not mention image-based sexual abuse content and instead tells users that this is “NOT the place to post, share, or promote... harmful or dangerous content” (TikTok, 2019; para 4).

On some platforms, the prohibition of image-based sexual abuse is unclear. For instance, Snapchat states that users should not “take Snaps of people in private spaces – like a bathroom, locker room or a medical facility – without their knowledge and consent” (Snap Inc., 2019, para 4). Although examples are given of what a “private space” might entail, it is unclear whether the nonconsensual sharing of nude or sexual imagery is also prohibited in the context of “public” spaces. Facebook’s policy on the sharing of image-based sexual abuse content, on the other hand, is much clearer, allowing the sharing of images to be either “noncommercial” or “private” with an expansive definition of what an “intimate” image includes. Facebook prohibits the nonconsensual sharing of intimate images according to three criteria: the image is noncommercial or produced in a private setting; the person is nude, nearly nude, or engaged in a sexual act or posing in a sexual way; and there is lack of consent indicated by captions, comments, the title of the page, independent sources, or reports from victims or others (Facebook, 2020a). However, the focus on images that are noncommercial, and which are produced in a private setting, appears to deny sex workers or pornographic actors the right to control the dissemination of their images.

There can be significant flow-on effects of ambiguous policy stances on image-based sexual abuse. Platform policies that are open-textured, or which use nondescript terms, can enable ad hoc decision-making in response to business and other pressures (Witt et al., 2019). The lack of consistent language for platforms to name and work through the problems of image-based sexual abuse can make it difficult for stakeholders to discuss the concerns that victim-survivors and other societal actors raise. Moreover, vague guidelines can fundamentally limit the ability of victim-survivors or their authorized representatives to apply platform policies to reporting features or inform users as to the bounds of acceptable behavior.

Given that platforms almost always reserve “broad discretion” to determine what, if any, response will be given to a report of harmful content (Suzor, 2019, p. 106), it is essentially their choice whether or not to impose punitive (or other) measures on users when their terms of service or community guidelines have been violated (some of which have appeals processes in place). While platforms are not able to make arrests or issue warrants, they are able to remove content, limit access to their sites to offending users, issue warnings, disable accounts for specified periods of time, or permanently suspend accounts at their discretion. YouTube, for instance, has implemented a “strikes system” which first entails the removal of content and a warning issued (sent by email) to let the user know the Community Guidelines have been violated with no penalty to the user’s channel if it is a first

offense (YouTube, 2020, What happens if, para 1). After a first offense, users will be issued a strike against their channel, and once they have received three strikes, their channel will be terminated. Other platforms have similar systems in place. As noted by York and Zuckerman (2019), the suspension of user accounts can act as a “strong disincentive” to post harmful content where social or professional reputation is at stake (p. 144).

Deepfakes

The extent to which platform policies and guidelines explicitly or implicitly cover “deepfakes,” including deepfake pornography, is a relatively new governance issue. Deepfakes are a portmanteau of “deep learning,” a subfield of narrow artificial intelligence (AI) used to create content and fake images. In December 2017, a Reddit user, who called himself “deepfakes,” trained algorithms to swap the faces of actors in pornography videos with the faces of well-known celebrities (see Chesney & Citron, 2019; Franks & Waldman, 2019). Since then, the volume of deepfake videos on the internet has increased exponentially; the vast majority of which are pornographic and disproportionately target women (Ajder, Patrini, Cavalli, & Cullen, 2019).

In early 2020, Facebook, Reddit, Twitter, and YouTube announced new or altered policies prohibiting deepfake content. In order for deepfake content to be removed on Facebook, for instance, it must meet two criteria: first, it must have been “edited or synthesized... in ways that aren’t apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say”; and second, it must be the product of AI or machine learning (Facebook, 2020a, Manipulated media, para 3). The narrow scope of these criteria, which appears to be targeting manipulated fake news rather than different types of manipulated media, makes it unclear whether videos with no sound will be covered by the policy – for instance, a person’s face that is superimposed onto another person’s body in a silent porn video. Moreover, this policy may not cover low-tech, non-AI techniques that are used to alter videos and photographs – also known as “shallowfakes” (see Bose, 2020).

On the other hand, Twitter’s new deepfake policy refers to “synthetic or manipulated media that are likely to cause harm” according to three key criteria: first, if the content is synthetic or manipulated; second, if the content was shared in a deceptive manner; and third, if the content is likely to impact public safety or cause serious harm (Twitter, 2020, para 1). The posting of deepfake imagery on Twitter can lead to a number of consequences depending on whether any or all of the three criteria are satisfied. These include applying a label to the content to make it clear that the content is fake; reducing the visibility of the content or preventing it from being recommended; providing a link to additional explanations or clarifications; removing the content; or suspending accounts where there have been repeated or severe violations of the policy (Twitter, 2020).

While specific deepfake policies do not exist on other platforms, some have more general rules relating to “fake/d,” “false,” “misleading,” “digitally manipulated,”

“lookalike,” and/or “aggregate” content, which could result in the take-down of deepfake images. Pornhub (2020) does not mention deepfakes in its Terms of Service; however, in 2018 it did announce a ban on deepfakes (Cole, 2018). Nevertheless, the site continues to host deepfake pornography. When we searched for “deepfakes” using the internal Pornhub search function, no results were found, yet when we searched through Google “deepfakes” and “pornhub,” multiple results of fake celebrity videos were returned.

Reporting Harmful Content

Reporting options are another means through which digital platforms can address the problem of image-based sexual abuse. All of the platforms we examined have in place some sort of reporting protocol, including the porn sites, which are supposed to trigger review by human content moderators. On porn sites, for instance, users can report through either a *Digital Millennium Copyright Act of 1998* take-down request, or via a content removal form. Facebook recently announced that image-based sexual abuse content is now triaged alongside self-harm in the content moderation queue (Solon, 2019).

Another important form of content reporting occurs through the “flagging” system where users are enlisted as a “volunteer corps of regulators” to alert platforms about content that violates their policies and community standards (Crawford & Gillespie, 2016, p. 412). Facebook users, for instance, flag around one million pieces of content per day (Buni & Chemaly, 2016). Many companies provide built-in reporting features through which users can report material that potentially violates content policies (Witt et al., 2019, p. 577). For instance, Pornhub allows users to flag videos (using the “Flag this video” link under each video) if it is “illegal, unlawful, harassing, harmful, offensive, or various other reasons,” stating that it will remove the content from the site without delay (Pornhub, 2020, Prohibited uses, para 2).

Platform reporting systems predominantly place the onus on victim-survivors or other users to flag or report image-based sexual abuse content. In other words, digital platforms “[responsibilize users] to reduce their own risk of [victimization]” (Salter, Crofts, & Lee, 2018, p. 301). Major online platforms, like Facebook and Instagram, suggest that users take a range of preventive measures, such as unfollowing or blocking those responsible for posting abusive content, reviewing their safety and security settings, and accessing hyperlinked information. Microsoft, for instance, suggests that users should identify the source and/or owner of an image and attempt to have it removed before reporting it as a potential policy violation (Microsoft, 2020). If unsuccessful, victims are encouraged to report content through built-in or other reporting features. Preconditions like this, in many ways, are a “practical solution to the problem of moderating vast amounts of content” (Witt et al., 2019, p. 576). However, while important safety and empowerment messages should be communicated to users, in isolation they can place additional emotional, financial, and other burdens on already vulnerable individuals.

Technological Tools

The third option for enhancing proactive platform action in relation to harmful content is to use technological solutions to prevent users perpetrating further abuse. In 2009, Microsoft and Professor Hany Farid from Dartmouth College developed PhotoDNA, a technology that creates a unique digital signature (also known as a “hash” or “digital fingerprint”) of child sexual abuse images, which can then be compared against known images stored in a database curated by the National Center for Missing and Exploited Children in the United States (Langston, 2018, para 13). This technology assists platforms to detect, remove, and block child sexual abuse content on their networks. It is also used by law enforcement to detect, arrest, and prosecute perpetrators, and identify victims.

PhotoDNA technology has led to other technological innovations in regulating harmful online content.¹ In November 2017, Facebook announced a pilot trial in partnership with the Australian Office of the eSafety Commissioner to prevent image-based sexual abuse from occurring on Facebook-owned platforms, which was then expanded to Canada, the United States, and the United Kingdom in May 2018 (Facebook, 2020b). The trial allows people who are concerned that someone might share an image of them to contact the relevant partner agency and complete an online form. The person is then sent an email containing a secure, one-time upload link, where she or he can upload the image. A community operations analyst from Facebook then accesses the image and creates a “hash” of it. If any user in the future attempts to upload or share that same image on the platform, they will be automatically blocked, and the image will not be able to be shared (Facebook, 2020b).

Other companies have adopted similar methods. Pornhub uses a third-party automated audio-visual identification system (called MediaWise®) which first identifies the content using “digital fingerprinting,” and then blocks it from being uploaded again in the future (Pornhub, n.d.). To have content digitally fingerprinted, victims are required to email a third-party service to make the request. The victim then receives an email to let them know the content has been fingerprinted and the victim can then report the video by filling out the online form on the site. It is important to note that fingerprinted videos that are blocked can still appear on the site, albeit for a brief time, which can have significant impacts on victim-survivors.

These technological solutions have received significant criticism. The Facebook pilot was widely condemned for asking vulnerable individuals to trust Facebook with their intimate images in the wake of the Cambridge Analytica scandal (see, e.g., Romano, 2018; see also Bailey & Liliefeldt, this volume). More recently, a Motherboard investigation found that Pornhub’s fingerprinting system “can be easily and quickly circumvented with minor editing” to change the metadata and therefore prevent the image being matched to the original image stored in the database (Cole & Maiberg, 2020, para 7). The Motherboard investigators, with the consent and cooperation of several women who were featured in nonconsensual pornographic videos, tested Pornhub’s content removal system, finding that Pornhub did remove the videos when they reported the content. The investigators

also tested the digital fingerprinting system by using editing techniques to alter the videos. They found that after the content had been flagged, removed, and fingerprinted, when they tried to upload the exact same video to Pornhub, the video was removed within an hour. However, they also experimented with using editing techniques to slightly alter a number of videos and found that when they did this, the fingerprinting method did not work and the video was not removed (Cole & Maiberg, 2020). Pornhub's system may be compared with Microsoft's PhotoDNA program where the hashed images are resistant to complex image alterations, including resizing and minor color alterations (Langston, 2018).

In 2019, Facebook introduced a new AI tool that can detect nonconsensually shared “nearly nude” images (Davis, 2019). Using a database of previously confirmed nonconsensual intimate images, the technology works by training the algorithm to recognize language patterns and key words that would suggest those images are not consensual (similar to Google's AI tool for detecting child abuse material). Once the content has been flagged, a member of Facebook's Community Operations team reviews the content and then decides whether the content has violated Facebook's community standards. If they conclude there has been a violation, Facebook will then disable the account or issue a warning to the user (Davis, 2019). This is an important proactive measure for two key reasons. First, it puts the onus back on the platform to find and remove image-based sexual abuse content. Second, it prevents the viewing and/or further sharing of these images. This is crucial since many victim-survivors will not know that images of them are being shared or will only discover their images well after they have been shared online. Nonetheless, it is important to note that this tool is imperfect because in many cases there will be no clear indication that the image has been shared without consent (e.g., clear indications include a victim makes a report, they are underage, or the accompanying text suggest vindictiveness). In cases where there is no clear indication, there is little Facebook and its AI tool can do to determine whether the image has in fact been taken or shared without consent.

Overall, these automated systems are revolutionary, and are helping some platforms to better address and prevent harmful online content. However, end-to-end encryption, where no one (including the platforms) can see the content of sent messages, works to circumvent the use of image hashing systems, thwarting global efforts to reduce the circulation of child sexual abuse or image-based sexual abuse material (see Green, 2019). This is a stark reminder that we cannot and should not rely on technological solutions alone to address the circulation of harmful content online.

Corporate Social Responsibility and Image-Based Sexual Abuse

Having outlined four main shortcomings associated with *governance* by platforms, in this final section we explore how CSR frameworks can provide critical guidance to digital platforms and help to set new norms to ensure that these governing actors take more proactive steps to both address and prevent the circulation of harmful content on their networks. CSR is an “essentially contested” term (Okoye, 2009) because

there is little agreement about what it entails and no universally agreed-upon definition (Ihlen, Bartlett, & May 2011). At its broadest, CSR is defined as “the business and society relationship” (Laidlaw, 2017, p. 138), or, at its narrowest, as “social responsibility” that “begins where the law ends” (Davis, 1973, p. 313). There are several justifications for greater social responsibility by platforms; chief among them is the role that these businesses play in facilitating, and in some instances even promoting, harmful content on their networks (Slane & Langlois, 2018, p. 46).

While there is little agreement about the nature and extent of corporations’ ethical or moral responsibilities (see Taddeo & Floridi, 2016), companies are increasingly being measured against a set of benchmarks on governance processes and respect for users’ privacy and freedom of expression. For instance, the Ranking Digital Rights’ (RDR) Corporate Accountability Index (2019) is an industry-level initiative that evaluates 24 of the world’s most powerful technology companies according to three key criteria: privacy; freedom of expression; and governance. Other global initiatives, such as the United Nations Guiding Principles (UNGPs) on Business and Human Rights, similarly provide a set of guidelines for transnational corporations, businesses, and states to prevent, address, and remedy human rights violations committed in business operations (see United Nations, 2011). There are, of course, several limitations associated with CSR-initiatives like these. In terms of RDR’s Index, only a small number of companies are ranked (none are porn sites) according to 35 broad indicators which measure performance. Unfortunately, they tell us little about how different companies are performing in relation to addressing specific online behaviors, such as image-based sexual abuse. Another concern is that most CSR frameworks, like the UNGPs, impose nonbinding – sometimes described as “blurred,” and “soft” – human rights obligations (see Jorgensen, 2017, p. 281; see also Coombs, this volume). Nonetheless, CSR initiatives have led to tangible improvements in the practices of technology companies relating to privacy, freedom of expression, and governance (RDR, 2019, p. 9).

We argue that CSR provides a useful lens through which to examine platform governance. Crane, Matten, and Spence’s (2013) conception of CSR comprises six main characteristics: (1) voluntary participation; (2) managing externalities such as impacts on local communities; (3) treating stakeholders as more than just shareholders; (4) aligning social and economic responsibilities; (5) forming business practices and values that address social issues; and (6) moving beyond just philanthropy. These characteristics not only underline what positive action by digital platforms arguably should look like, but they also provide a useful language to identify and work through potential issues. We draw on this conception of CSR in our discussion below by focusing on four key barriers to addressing and preventing image-based sexual abuse on digital platforms. Specifically, we demonstrate the ways in which *governance by* platforms frequently conflicts with these CSR ideals.

Barriers to Corporate Social Responsibility

First, there is reluctance on the part of some companies to voluntarily or proactively intervene to address the problem of harmful online content. This reluctance is understandable from a strict legal perspective when considering the distinction between the responsibilities of digital platforms and the duties and authority of

nation states. As mentioned above, CDA 230, for example, does not (with some exceptions) require digital platforms to moderate content (Tushnet, 2008, pp. 1001–1002). For instance, Pornhub's (2020) Terms of Service make clear that they do not have an obligation to review content and do not regularly do so.² There are also inherent difficulties in weighing the potentially conflicting demands between freedom of speech and other rights, which form the foundation of some arguments against corporate regulation of content (see Hintz, 2014, p. 349). Moreover, there are well-founded concerns about vesting largely nontransparent and unaccountable regulatory power in private entities, some of which may make mistakes when moderating content by prioritizing commercial interests over ethical issues and community-oriented or social justice goals.

A second related barrier to CSR is economic profit. Platforms are generally focused on attracting more users and thereby generating greater advertising revenue, which can have far-reaching impacts on users' data and privacy. Digital platforms, however, do regularly engage in philanthropic, socially responsible activity, seeking to ostensibly align economic profit with social responsibility. By way of example, in March 2020, in response to the devastation wrought by COVID-19, Pornhub donated surgical masks and some of its proceeds to sex workers affected by the pandemic (Iovine, 2020).

These philanthropic acts stand in stark contrast to a range of nonethical and criminal practices on porn sites like Pornhub. In March 2020, nearly half a million people signed an online petition (started by a group called Exodus Cry) to hold Pornhub accountable for hosting child rape and underage porn videos, some of whom are victims of sex trafficking (Milne, 2020). The petition also claims that Pornhub does not have a reliable system in place to verify the age or consent of those featured in the pornographic content on the site. One key challenge is that on porn sites, nonconsensual content is an extremely profitable enterprise, generating millions of views and attracting millions of dollars of advertising. As such, there may be very little incentive for porn companies to remove image-based sexual abuse content, and little risk that failing to do so will damage their corporate image. Mainstream social media platforms, on the other hand, may have more incentive to address harmful online content because failing to do so could result in significant damage to their corporate reputation.

A third barrier relates to the significant technical, logistical, and emotional challenges that come with attempting to regulate massive volumes of online content (Laidlaw, 2017; Roberts, 2019). This is particularly so on platforms such as Facebook which currently has 2.5 billion active users (Hutchinson, 2020). The sheer scale of content means that not only are there delays in content review and removal processes, but content moderators – many of whom are contracted workers on low-pay – are regularly exposed to violent and harmful content causing some to experience significant vicarious trauma (see Boran, 2020).

Finally, owing to these aforementioned barriers, there is often a gap between the policies on image-based sexual abuse and the actual practice of content review and removal. US lawyer Carrie Goldberg (2019) has commented on Google's lack of action in taking down videos of women who were deceived into performing in porn or sexually assaulted in those videos. According to Goldberg: "If Google decides it

will keep linking to a website that contains your nude images, victims are just out of luck. And there's no appellate body. There is no law, only corporate policy, that protects (or fails to protect) victims' most private information" (para 9). Overall, little remains known about the number of images flagged by users, reported by victim-survivors, tagged for review by AI systems, or fingerprinted for future blocking. This in turn makes it difficult for stakeholders to determine how effective platform governance has been in practice for responding to and preventing image-based sexual abuse.

Some technology companies have faced public scrutiny for their opaque content moderation processes (see [Hopkins, 2017](#)), leading in some instances to greater transparency by these companies. For instance, Facebook's Transparency Report is a regular reporting system that gives "community visibility" to how Facebook enforces its community standards, protects intellectual property, responds to legal requests for user data or content restrictions, and monitors internet disruptions across its products ([Facebook, 2020c](#), para 1). Facebook also has a companion guide that explains how they write their policies, how they find and review potential violations, and how they measure results ([Facebook, 2020d](#)). In relation to the enforcement of its Community Standards, Facebook publicizes metrics on how they are "preventing and taking action on content that goes against these policies" in relation to a number of different issues, such as "adult nudity and sexual activity," "bullying and harassment," "child nudity and sexual exploitation of children," and hate speech ([Facebook, 2020c](#), Community standards enforcement report, para 1 & 4). Regarding "adult nudity and sexual activity," Facebook reports metrics on the prevalence of this content, how much content Facebook took action on, how much content was found by Facebook before it was reported by users, how much content was appealed by users, and how much actioned content was later restored by Facebook ([Facebook, 2020c](#), Community standards enforcement report, para 4). Unfortunately, the Facebook report does not break down the different types of adult nudity and sexual activity – which includes depiction of nudity as well as image-based sexual abuse content. In other words, Facebook does not report specifically on the prevalence of image-based sexual abuse content on their platform nor the actions that Facebook took to remove this content or the consequences for violating users. Nevertheless, Facebook's attempts at greater transparency in relation to content moderation are an example of more accountable forms of CSR, which many other companies have not yet adopted.

Given the barriers to addressing and preventing image-based sexual abuse through CSR-initiatives, as well as the four main shortcomings that we identified with the governance practices of our selected platforms, we argue that governments can and should play a greater role in addressing deficiencies in *governance by platforms*. This is not to suggest that platforms should be held to the same standards as nation states. Top-down, command, and control regulatory responses are generally unsuitable for decentered contexts like the internet and not necessarily desirable for users, lawmakers, and other stakeholders ([Witt et al., 2019](#), p. 593). Rather, we argue what is needed is more state-based regulation (*governance of platforms*) that can help to better align *governance by platforms* with CSR – specifically, initiatives embedded within social justice and human rights

frameworks. One positive step forward would be an amendment to CDA 230 so that the safe harbor provision only applies to those platforms that take reasonable steps to review and remove harmful content (see [Citron, 2018](#)).

The suggestion to introduce more government regulation is not, in our view, radical. As we have shown in the first section of this chapter, the regulatory landscape around the *governance of* platforms is changing, with many governments around the globe taking steps to hold technology companies responsible for hosting harmful content ([Flew et al., 2019](#)). Indeed, in early 2019, Facebook Co-Founder and CEO Mark Zuckerberg acknowledged: “Lawmakers often tell me we have too much power over speech, and frankly I agree. I’ve come to believe that we shouldn’t make so many important decisions about speech on our own” ([Zuckerberg, 2019](#), para 5). We stress, however, that any attempt to regulate platforms must be reasonable, proportionate, and take into account guides to good regulatory design, such as the *Manila Principles on Intermediary Liability (2015)* which seeks to balance user rights to freedom of expression, freedom of association, and the right to privacy. These steps would further propel a new culture of digital platform governance that is predicated not solely on economic profit, but on social justice, community, and ethics.

Conclusion

This chapter has examined the problem of image-based sexual abuse in the context of both the *governance of* and *by* online platforms. In terms of the *governance of* platforms, we paid particular attention to platform protection under CDA 230, which plays an important role in enabling platforms to govern in “lawless” ways ([Suzor, 2019](#)). One of the main ways that platforms govern their networks is by moderating user-generated content ([Witt et al., 2019](#), p. 557). Moderation processes, including platform-specific rules around content and the online architecture of platforms themselves, can make it easier or harder for users to undertake certain types of behavior ([Suzor, 2019](#), p. 91). In more recent years, other measures such as reporting options, digital fingerprinting, and other automated detection systems have become an essential part of the repertoire for tackling image-based sexual abuse.

Despite good intentions and significant changes to governance by platforms in recent years, we found that the policies, tools, and practices that are designed to address and prevent image-based sexual abuse are often piecemeal and reactive. Specifically, we identified four main issues: (1) inconsistent, reductionist, and ambiguous language in content policies; (2) a stark gap between the policy and practice of moderating content, including significant transparency deficits; (3) imperfect technology for detecting abuse, given that even where content is removed, images can still appear again on those sites or can be easily circulated on other platforms; and (4) the onus continues to remain predominantly with victim-survivors to report and prevent abuse. Overall, we argued that when platforms fail to address these issues, they risk failing victim-survivors and are implicated in the perpetration

of image-based sexual abuse. In response, we called for state-based regulation that can help to better align *governance by* platforms with CSR initiatives.

There are a number of steps that platforms themselves can take to better address the scourge of image-based sexual abuse on their networks: principally, adopting a multifaceted, community-oriented, and social justice-based regulatory approach. Such an approach should include clear and robust policies that specifically prohibit image-based sexual abuse content, with punitive and educative functions clearly attached; architectural modifications, including better systems to reliably verify both the age and consent of those featured in content hosted on platforms; more resources invested into AI and other automated systems for detecting and removing content, with careful attention paid to the shortcomings of these technological solutions; greater multi-stakeholder collaboration and consultation between digital platforms with civil society and government actors to achieve common objectives; and increased public debate about the *governance of* platforms and what CSR should entail. Most importantly, *governance by* platforms should be transparent and accountable, subject to the scrutiny of civil society, and flexible in an ever-changing digital landscape.

Notes

1. The detection of deepfake imagery using machine learning is another technological challenge that requires vigilant testing and refinement to keep up with the rapidly changing development of deepfake technology.
2. Pornhub (2020) also states that it is not responsible for any links to third-party websites that are not owned or controlled by Pornhub, and that it “will not and cannot censor or edit the content of any third-party site” (About our websites, para 2). Indeed, the Pornhub Terms of Service, ad nauseum, claims no liability “for any action or inaction regarding transmissions, communications or Content provided by any user or third party” (Pornhub, 2020, Monitoring and enforcement, para 5). Their Terms of Service also make it clear that users are “solely responsible” for the content and the consequences of posting content (Pornhub, 2020).

References

- Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019, September). The state of deep-fakes: Landscape, threats, and impact. Retrieved from <https://deeptracelabs.com/mapping-the-deepfake-landscape/>
- Alexa Internet. (n.d.). The top 500 sites on the web. Retrieved from <https://www.alexa.com/topsites>
- Boran, M. (2020, February 27). Life as a Facebook moderator: People are awful. This is what my job has taught me. *The Irish Times*. Retrieved from <https://www.irishtimes.com/>
- Bose, N. (2020, January 9). U.S. lawmakers say Facebook’s steps to tackle ‘deepfake’ videos are not adequate. *Reuters*. Retrieved from <https://www.reuters.com/>
- Buni, C., & Chemaly, S. (2016, April 13). The secret rules of the internet. *The Verge*. Retrieved from <https://www.theverge.com>

- Chesney, D., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1819. doi: [10.15779/Z38RV0D15J](https://doi.org/10.15779/Z38RV0D15J)
- Citron, D. (2018). Section 230's challenge to civil rights and civil liberties. (Legal Studies Research Paper No. 2018–18). Retrieved from <https://ssrn.com/abstract=3193214>
- Citron, D., & Franks, M. (2014). Criminalizing revenge porn. *Wake Forest Law Review*, 49, 345–391. doi: [10.1177/1557085117698752](https://doi.org/10.1177/1557085117698752)
- Cole, S. (2018, February 6). Pornhub is banning AI-generated fake porn videos, says they're nonconsensual. *Vice*. Retrieved from <https://www.vice.com>
- Cole, S., & Maiberg, E. (2020, February 7). Pornhub doesn't care. *Vice*. Retrieved from <https://www.vice.com>
- Crane, A., Matten, D., & Spence, L. J. (2013). Corporate social responsibility in a global context. In A. Crane, D. Matten, & L. J. Spence (Eds.), *Corporate social responsibility: Readings and cases in a global context* (2nd ed., pp. 3–26). Oxon: Routledge.
- Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), 410–428. doi: [10.1177/1461444814543163](https://doi.org/10.1177/1461444814543163)
- Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019, No. 38. Retrieved from <https://www.legislation.gov.au/Details/C2019A00038>
- Davis, K. (1973). The case for and against business assumption of social responsibilities. *Academy of Management Journal*, 16(2), 312–322. doi: [10.2307/255331](https://doi.org/10.2307/255331)
- Davis, A. (2019, March 15). Detecting non-consensual intimate images and supporting victims. *Facebook*. Retrieved from <https://about.fb.com/news/2019/03/detecting-non-consensual-intimate-images/>
- Echikson, W., & Knodt, O. (2018, November). Germany's NetzDG: A key test for combatting online hate. (Report No. 2018/09). Retrieved from http://wp.ceps.eu/wp-content/uploads/2018/11/RR%20No2018-09_Germany's%20NetzDG.pdf
- The European Parliament and the Council of the European Union. (2016/679). General Data Protection Regulation. Retrieved from <https://gdpr-info.eu/>
- Facebook. (2020a). Community standards. Retrieved from <https://www.facebook.com/communitystandards/introduction>
- Facebook. (2020b). The pilot. Retrieved from <https://www.facebook.com/safety/not-withoutmyconsent/pilot>
- Facebook. (2020c). Facebook Transparency Report. Retrieved from <https://transparency.facebook.com/>
- Facebook. (2020d). Understanding the community standards enforcement report. Retrieved from <https://transparency.facebook.com/community-standards-enforcement/guide>
- Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, 10(1), 33–50. doi: [10.1386/jdmp.10.1.33_1](https://doi.org/10.1386/jdmp.10.1.33_1)
- Franks, M. A., & Waldman, A. E. (2019). Sex, lies, and videotape: Deep fakes and free speech delusions. *Maryland Law Review*, 78(4), 892–898. Retrieved from <https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=3835&context=mlr>
- Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, CT: Yale University Press.
- Goldberg, C. (2019, August 17). How Google has destroyed the lives of revenge porn victims. *New York Post*. Retrieved from <https://nypost.com>

- Green, M. (2019). Can end-to-end encrypted systems detect child sexual abuse imagery?. [Blog post]. Retrieved from <https://blog.cryptographyengineering.com/2019/12/08/on-client-side-media-scanning/>
- Henry, N., & Flynn, A. (2019). Image-based sexual abuse: Online distribution channels and illicit communities of support. *Violence Against Women*, 25(16), 1932–1955. doi: [10.1177/1077801219863881](https://doi.org/10.1177/1077801219863881)
- Henry, N., McGlynn, C., Flynn, A., Johnson, K., Powell, A., & Scott, A. J. (2020). *Image-based sexual abuse: A study on the causes and consequences of non-consensual nude or sexual imagery*. London and New York, NY: Routledge.
- Hintz, A. (2014). Outsourcing surveillance—privatising policy: Communications regulation by commercial intermediaries. *Birbeck Law Review*, 2(2), 349–368. Retrieved from <http://www.bbkrlr.org/2-2-10.html>
- Hopkins, N. (2017, May 22). Revealed: Facebook’s internal rulebook on sex, terrorism and violence. *The Guardian*. Retrieved from <http://www.theguardian.com>
- Hutchinson, A. (2020, January 30). Facebook climbs to 2.5 billion monthly active users, but rising costs impede income growth. *Social Media Today*. Retrieved from <https://www.socialmediatoday.com>
- Ihlen, Ø., Bartlett, J., & May, S. (Eds.), (2011). *The handbook of communication and corporate social responsibility*. West Sussex: Wiley-Blackwell.
- Innovation, Science and Economic Development Canada. (2019). *Canada’s Digital Charter in action: A plan by Canadians for Canadians*. (Catalogue no. Iu4-259/2019E-PDF). Ottawa, ON: Government of Canada. Retrieved from [https://www.ic.gc.ca/eic/site/062.nsf/vwapj/Digitalcharter_Report_EN.pdf/\\$file/Digitalcharter_Report_EN.pdf](https://www.ic.gc.ca/eic/site/062.nsf/vwapj/Digitalcharter_Report_EN.pdf/$file/Digitalcharter_Report_EN.pdf)
- Iovine, A. (2020, March 25). Pornhub Premium is now free for everyone to encourage you to stay home. *Mashable*. Retrieved from <https://mashable.com>
- Jørgensen, R. F. (2017). What platforms means when they talk about human rights. *Policy & Internet*, 9(3), 280–296. doi:[10.1002/poi3.152](https://doi.org/10.1002/poi3.152)
- Kemp, S. (2020, January 30). Digital 2020: 3.8 billion people use social media. [Blog post]. Retrieved from <https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media>
- Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131, 1598–1670. Retrieved from https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf
- Laidlaw, E. (2017). Myth or promise? The corporate social responsibilities of online service providers for human rights. In M. Taddeo & L. Floridi (Eds.), *The responsibilities of online service providers* (pp. 135–154). Cham: Springer International Publishing.
- Langston, J. (2018, September 12). How PhotoDNA for video is being used to fight online child exploitation. *Microsoft*. Retrieved from <https://news.microsoft.com>
- Lee, K. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Boston, NY: Houghton Mifflin Harcourt.
- Manila Principles on Intermediary Liability. (2015). Retrieved from <https://www.manilaprinciples.org/>
- Martens, T. (2011, December 4). Rockers, fully exposed on Is Anyone Up?. *Los Angeles Times*. Retrieved from <https://www.latimes.com/entertainment/la-xpm-2011-dec-04-la-ca-pop-nudes-20111204-story.html>

- McGlynn, C., & Rackley, E. (2017). Image-based sexual abuse. *Oxford Journal of Legal Studies*, 37(3), 535–561. doi:10.1093/ojls/gqw033
- Microsoft. (2020). Non-consensual pornography reporting form. Retrieved from <https://www.microsoft.com/en-au/concern/revengeporn>
- Milne, A. (2020, March 27). Porn site's free service during coronavirus raises sex trafficking fears. *Reuters*. Retrieved from <https://www.reuters.com/article/us-britain-women-trafficking-trfn/porn-sites-free-service-during-coronavirus-raises-sex-trafficking-fears-idUSKBN21D3E9>
- Okoye, A. (2009). Theorising corporate social responsibility as an essentially contested concept: Is a definition necessary?. *Journal of Business Ethics*, 89, 613–627. doi:10.1007/s10551-008-0021-9
- Pornhub. (2020, March 9). Terms of service. Retrieved from <https://www.pornhub.com/information#terms>
- Pornhub. (n.d.). Content removal request?. Retrieved from <https://www.pornhub.com/content-removal>
- Powell, A., Henry, N., & Flynn, A. (2018). Image-based sexual abuse. In W. DeKeseredy, & M. Dragiewicz (Eds.), *Routledge handbook of critical criminology* (2nd ed., pp. 305–315). Abingdon and New York, NY: Routledge.
- Ranking Digital Rights. (2019, May). 2019 Corporate Accountability Index. Retrieved from <https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf>
- Roberts, S. (2019). *Behind the screen: Content moderation in the shadows of social media*. New Haven, CT: Yale University Press.
- Romano, A. (2018, May 24). Facebook's plan to stop revenge porn may be been creepier than revenge porn. *Vox*. Retrieved from <https://www.vox.com/>
- Ruvalcaba, Y., & Eaton, A. A. (2020). Nonconsensual pornography among US adults: A sexual scripts framework on victimization, perpetration, and health correlates for women and men. *Psychology of Violence*, 10(1), 68–78. doi:10.1037/vio0000233
- Salter, M., Crofts, T., & Lee, M. (2018). Beyond criminalisation and responsabilisation: Sexting, gender and young people. *Current Issues in Criminal Justice*, 24(3), 301–316. doi:10.1080/10345329.2013.12035963
- Satariano, A. (2018, May 24). G. D. P. R., a new privacy law, makes Europe world's leading tech watchdog. *New York Times*. Retrieved from <https://www.nytimes.com>
- Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department. (2019). Online Harms White Paper. London: HM Government. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf
- Slane, A., & Langlois, G. (2018, March). Debunking the myth of “not my bad”: Sexual images, consent, and online host responsibilities in Canada. *Canadian Journal of Women and the Law*, 30(1), 42–81. doi:10.3138/cjwl.30.1.42
- Snap Inc. (2019). Community guidelines. Retrieved from <https://www.snap.com/en-US/community-guidelines>
- Solon, O. (2019, November 18). Inside Facebook's efforts to stop revenge porn before it spreads. *NBC News*. Retrieved from <https://www.nbcnews.com>
- Suzor, N. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge: Cambridge University Press.

- Taddeo, M., & Floridi, L. (2016). The debate on the moral responsibilities of online service providers. *Science and Engineering Ethics*, 22, 1575–1603. doi:10.1007/s11948-015-9734-1
- TikTok. (2019, December 12). Understanding our community guidelines. Retrieved from <https://newsroom.tiktok.com/understanding-our-community-guidelines/>
- Tumblr. (2020, January 23). Community guidelines. Retrieved from <https://www.tumblr.com/policy/en/community>
- Tushnet, R. (2008). Power without responsibility: Intermediaries and the First Amendment. *George Washington Law Review*, 76(4), 986–1016. Retrieved from <https://www.gwlr.org/wp-content/uploads/2012/08/76-4-Tushnet.pdf>
- Twitter. (2020). Synthetic and manipulated media policy. Retrieved from <https://help.twitter.com/en/rules-and-policies/manipulated-media>
- United Nations. (2011). Guiding principles on business and human rights. Retrieved from https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf
- Westlake, E. J. (2008). Friend me if you Facebook: Generation Y and performative surveillance. *TDR/The Drama Review*, 52(4), 21–40. doi:10.1162/dram.2008.52.4.21
- Witt, A., Suzor, N., & Huggins, A. (2019). The rule of law on Instagram: An evaluation of the moderation of images depicting women's bodies. *UNSW Law Journal*, 42(2), 557–596. Retrieved from <http://www.unswlawjournal.unsw.edu.au/wp-content/uploads/2019/06/6-UNSWLJ-422-Witt-Suzor-and-Huggins-Final.pdf>
- xHamster. (2020, April 22). Terms and conditions: User agreement. Retrieved from <https://xhamster.com/info/terms>
- XVideos. (n.d). XVideos terms of service. Retrieved from <https://info.xvideos.com/legal/tos/>
- York, J. C., & Zuckerman, E. (2019). Moderating the public sphere. In R. F. Jørgensen (Ed.), *Human rights in the age of platforms* (pp. 137–161). Cambridge, MA: The MIT Press.
- YouTube. (2020). Nudity and sexual content policies. Retrieved from <https://support.google.com/youtube/answer/2802002?hl=en>
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. London: Profile Books.
- Zuckerberg, M. (2019, March 30). Mark Zuckerberg: The internet needs new rules. Let's start in these four areas. *Washington Post*. Retrieved from <https://washingtonpost.com>