Moderating online child sexual abuse material (CSAM): Does self-regulation work, or is greater state regulation needed?

Abstract

Social media platforms are crucial public forums connecting users around the world through a decentralised cyberspace. These platforms host high volumes of content and, as such, employ content moderators (CMs) to safeguard users against harmful content like child sexual abuse material (CSAM). These roles are critical in the social media landscape, however CMs' work as "digital first responders" is complicated by legal and systemic debates over whether the policing of cyberspace should be left to the self-regulation of technology companies, or if greater state-regulation is required. In this empirical policy and literature review, major debates in the area of content moderation and, in particular, the online policing of CSAM are identified and evaluated. This includes the issue of territorial jurisdiction, and how it obstructs traditional policing; concerns over free speech and privacy if CMs are given greater powers and; debates over whether technology companies should be legally liable for user-generated content (UGC). In outlining these issues, a more comprehensive foundation for evaluating current practices for monitoring and combatting online CSAM is established which illustrates both the practical and philosophical challenges of the existing status quo, wherein the state and private companies share these important responsibilities.

Keywords*:* content moderation; social media; online harms; child sexual abuse material; cybercrime; online policing.

Introduction

The Internet is a complex space for law enforcement to operate. It is a space where traditional 'bricks and mortar' models of policing and concepts like territorial jurisdiction are complicated by the nebulous, intangibility of a globally-connected world. In the past, these challenges have largely been subsumed by an overarching ethos of liberalism and self-regulation in cyberspace, with society relying on technology companies to voluntarily practice corporate social responsibility (CSR) by "moderating" the experience of users visiting their sites — essentially, deciding which UGC is appropriate to be posted and which should be removed, typically because it is either illegal or breaches the platform's own terms of service (Roberts, 2014). As social media platforms have increasingly assumed its role of the predominant space for public discourse in the modern age, the volume of UGC requiring such moderation has also increased. In turn, private content moderators (CMs) employed by social media platforms have effectively become the new "frontline" in the fight against harmful online content, including the proliferation of child sexual abuse material (CSAM).

The result is a somewhat uneasy coproduction of justice on the Internet, wherein state-based actors like traditional law enforcement have limited power and private actors like social media companies have a CSR as the only entities with total jurisdiction over what happens on the platforms they manage (Perloff-Giles, 2018; Nolte & Westermeier, 2020). Characterised by its preference for self-regulation, the traditional status quo has been a predominantly non-interventionist approach toward social media platforms from the international community. As the volume of harmful content shared on social media shows no signs of abating, questions continue to be raised as to whether or not this negotiated coproduction of online policing is actually sustainable in the long-term (Cusumano et al., 2021; Bischoff, 2022). This article aims

to explore some of the key challenges facing the policing of CSAM on social media as experienced by *both* parties involved in this coproduced response: state law enforcement agencies and their partners, the technology companies.

Based on a scoping review of existing literature, as well as additional socio-legal materials, several specific areas were identified where complications in the current system have either already arisen, or may in the future. These subtopics were selected as foundational concerns among a myriad of issues identified in the scoping review overall, with action on these particular concerns providing a basis for strengthening the system overall. This includes intrinsic jurisdictional issues posed by attempting to apply a traditional, physical policing approach to an online world where communications (and crimes) traverse international borders. These matters complicate policing efforts, and are central to the argument that self-regulation by social media companies is the only logical response to harmful online content. The paper also explores the impact that ideals of "net neutrality" and free speech have on social media platforms' ability (and willingness) to moderate content. This is closely tied to recent debates around the extent to which governments should intervene in the content moderation process. As campaigners in the United States argue for reforms to Section 230 of the *Communications Decency Act*, which would make social media companies more liable for content (such as CSAM) posted on their platforms, advocates for self-regulation make a case that the principles of self-regulation and CSR are sufficient, and that further government intervention would be inappropriate, or counterproductive. Finally, this article explores the work of CMs and the context in which it is conducted, considering the current system in terms of its transparency and its actual efficacy.

Literature Review

One crime that has benefited greatly from the interconnected structure and looser regulation of the Internet is the sharing of CSAM. Despite seeming straightforward on a superficial level, definitions of what CSAM is differ based on jurisdiction, as well as *who* is doing the defining. These differences can emerge from varied legal definitions of what constitutes a child (including disparities in the age of majority, or sexual consent, between countries), and distinctions as to whether or not non-visual content such as grooming is treated as CSAM (Johansson, 2019). For the purposes of this article, the definition used by the Council of Europe has been adopted, with CSAM defined as *any* sexual audio-visual depiction of children, ranging from children being 'posed' in a sexual manner to real depictions of sexual abuse (Council of Europe, 1993). This is an area of online offending that has seen consistent, exponential growth: Bursztein et al (2019) state that reports of online CSAM to United States authorities have increased by 50 percent each year for the last two decades. Despite this, Salter and Hanson (2021) describe the response of regulatory authorities in the United States, where the majority of technology companies are based, as "notoriously lax and oriented toward the growth and profitability of technology companies rather than child protection" (p. 731). They attribute this weaker governmental approach, in part, to "a pervasive cyberlibertarianism [which] played a major role in legitimizing anti-regulation ethos within industry and government" (Salter & Hanson, 2021, p. 731).

Previously, Salter (2017) observed that this anti-regulation ethos was exacerbated by the rise of social media platforms as dominant players in the digital space — with a business-model predicated on ever-increasing interactions and growing a user-base, he argued that social media platforms adopted a "hands-off" approach of self-declared neutrality that distanced them from

the substance of the content being posted on their platforms. To define *social media*, we draw on the literature-informed conceptualization provided by Obar and Wildman, which broadly treats social media as apps where (1) "user-generated content [UGC] is the lifeblood"; (2) individuals and/or groups create specific profiles or accounts, and; (3) platforms facilitate contact between users, allowing for the formation of social networks (2015, p. 745). This definition would include both mainstream public sites like Twitter and Facebook, as well as more "closed" platforms such as WhatsApp. The precarious balance between CSR to provide a safe online environment and this anti-regulation (or, *self-regulation*) ethos has been discussed by Edwards et al. (2021). In their research into the prevailing views in the social media industry and associated fields (e.g., academic, legal, and institutional entities), Edwards et al. reported an overarching sentiment that traditional enforcement measures were limited in the online space and, thus, stimulating self-regulation was essential; however, they also reported the key forecast that "the likelihood is that harm to vulnerable groups will be 'accommodated' in liberal democracies as a price to be paid for the perceived political and economic benefits of unmoderated social media" (2021, p. 1). Concerns about self-regulation in the social media industry are shared by Common (2020), who notes that "the problem with self-regulation is it suffers from a low degree of compliance … even when platforms make their own rules or voluntarily accede to codes, they do not necessarily adhere to them" (p. 145). Síthigh (2017) goes further, arguing that self-regulation is underpinned by "an over-emphasis on non-interventionist techniques … [which gives social media platforms] unintentionally significant power in violation of the communicative rights of individual users" (p. 86).

Despite these critiques, there is an argument to be made that self-regulation by social media platforms is the preferable model for managing CSAM. Common (2020) highlights a connection between self-regulation and the principles of CSR, a practice based on making a

"business case for human rights … [which] typically relies on corporate voluntarism" (p. 13). The business case for CSR and self-regulation is expressed by Gunningham and Rees (1997), who argue that self-regulation is more efficient and cheaper than government intervention for both companies and consumers, via costs passed on. Cusumano, Gawer and Yoffie (2021) build on this point, stating that "in effect, self-regulation usually shifts discretion over how to regulate from the regulator [the government] to the target [the company] … this can be beneficial because targets are likely to have greater knowledge of and more information about their own operation" (p. 1264). Cusumano, Gawer and Yoffie (2021) comment that the threat of more intrusive government intervention has been central to improved self-regulation efforts by social media companies in recent years, claiming that technology companies "now see [government] regulation as a real possibility, so they are more seriously beginning to self-regulate and propose innovative solutions" (p. 1282). The benefits of self-regulation are not only to private corporations: for governments, self-regulation in the online sector offers "the possibility of outsourcing enforcement and minimising the accompanying costs, while industry [is] attracted by the promise of a flexible and light touch regulatory regime" (McIntyre, 2013, p. 278). A status quo results from social media self-regulation which satiates both government and industry on a theoretical level; however, it is the successful, practical management and policing of CSAM that remains largely unresolved in the current context, and is the focus of this article's discussion.

Much of the practical side of policing harmful online content falls to commercial CMs, employed by technology companies in pursuit of their corporate social responsibility obligations. Referring to an earlier state of "cyberanarchy", Roberts (2014) argues that "the Internet is, in fact, predicated on control at every level … [Content moderation] practices fit into this cycle in such a way as to be undetected by most, and thus help constitute an illusion

of volition and participation that is not reality-based" (p. 68). Roberts goes on to note that, considering that *de*regulation (rather than increased regulation) has been a feature of the digital world for most of the Internet's history, users "have no recourse but to put their faith in the benevolence of private corporations over which they can exercise virtually no control" to ensure a safe and legal online world (2014, p. 85). By performing this role, Roberts (2016) asserts that CMs are "indispensable to the sites for which they labor … they curate site content and guard against serious infractions contained in UGC that might do harm to a social media platform or company's digital presence" (p. 147). As Steiger (2020) rightly notes, the role goes beyond "merely selecting 'delete'" wherever harmful content like CSAM is identified: instead, a CM's job "requires informed knowledge of social and cultural norms, [and] government regulations … along with the rigorous demands of ensuring the removal of inappropriate material according to platform and governance, workers must also meet stringent accuracy and efficiency scores" (pp. 3-4).

Rather than treating CMs as an abdication of securitisation to private entities, Bellanova and De Goede (2021) construct the process as one of "co-production between private expertise and public security … when security decisions – of referral, removal, flagging and filtering – are produced at the intersection between public policing and private platforms" (p. 5). Nolte and Westermeier (2020) concur that, in this context, "the public and the private are not two realms that can be analysed apart from each other" (p. 63). Langvardt (2018) puts the onus for 'patrolling' the Internet on social media companies. He asserts that "the Internet makes it easy for bad actors … [and] something must be done to mitigate this problem … [to do so] however, is to adopt a pervasive system of prior restraints based on snap judgements" — as Langvardt observes, this is a process that is currently managed almost exclusively by private CMs (pp. 1358-1359). In the past, the criminal justice system could afford to move at a slower pace in

response to crimes occurring via conventional 'old media' platforms. Langvardt says that "online platforms, by contrast, must move aggressively and quickly to suppress an enormous volume of unwanted communications" (2018, p. 1360). Scholars such as Banko, MacKeen and Ray (2020) have attempted to develop a "unified typology of harmful content" to guide content moderation, and social media platforms themselves promote the utility of algorithms designed to automate content moderation (e.g. Gillespie, 2020). However, the fact remains that the bulk of this process is at the discretion of CMs, who adopt the role of 'online police' in this private-led effort to securitise platforms and deal with CSAM.

Current practices on social media: a background

Data released by the technology industry and published in the grey literature reveals the scale of social media, and its importance as a primary space for human interaction in the modern world: a report from 2019 showed that 2.95 billion people globally utilized social media, with that figure estimated to rise to 3.43 billion by 2023, representing around half the global population (Clement, 2019). While, as Steiger notes, "most information read and shared by the average users generally appears to be benign by social standards" (2020, p.2), the total number of social media users represents not just those with potential to share harmful content, but the scope of how many people may be *exposed* to it. In a business transparency report released by Facebook in early 2022, the organisation revealed it had flagged 77.4 million posts for removal as "child nudity and sexual exploitation" in 2021; this represents around 40 million more posts than the total in this category for the entirety of 2020 (Meta, 2022). Other popular social media platforms experienced proportionally-similar levels of CSAM content-sharing in 2021. In the first three-quarters of 2021, Instagram removed around 4.8 million posts, TikTok received 33.7 million reports, and YouTube removed around 9 million videos; SnapChat removed 119,134 accounts for violations in the first six months of the year alone (Bischoff, 2022). While based

on different metrics, the data nevertheless illustrates the scale of illicit content-sharing on mainstream platforms and, importantly, that the rates of harmful content being circulated on these platforms continues to rise. In every case, the 2021 figures for reported and/or removed CSAM was considerably higher than the year prior, indicating that the issue of content moderation and management continues to grow, and place increasing pressure on those responsible for carrying it out.

For the most part, ensuring content posted on social media sites meets legal (and community) standards falls, in practice, to private CMs employed by the technology companies themselves. Depending on the protocols of the specific organisation, CMs may either proactively seek for potential breaches of their employers "community guidelines" or, more commonly, respond to reports of inappropriate content (Drootin, 2021). These reports often come directly to the social media platform from other users; however, technology companies have also adopted other methods to identify potentially harmful material for removal. One such avenue is the global INHOPE network, which operates "hotlines" in various countries where members of the public can report CSAM. Operating fifty hotlines worldwide, INHOPE analysts review all reports of CSAM, and classifies that material, entering it into an international database for further action (Dabrowska, 2021). Another means of managing content on social media is through blocking it on a macrolevel, via Internet Service Providers (ISPs) rather than social media platforms themselves. As a result of the United Kingdom's (much-replicated) self-regulatory model, pressure placed on ISPs by government led to the implementation of "Cleanfeed" systems designed to block access to harmful content; as a collective, the ISPs also established the Internet Watch Foundation (IWF) which employs analysts to respond to reported harmful content, determine if it is illegal under UK law and, if so, pass it on to ISPs to remove or block, and police for further action (Brown, 2010). Regardless of the specific mechanism, the

common factor in these systems is that they are primarily operated as a partnership between non-governmental organisations and private companies, without the direct involvement of the state. This, again, reflects the traditional preference for self-regulation over government intervention in this area, and reinforces the concept that social media CMs have assumed the position of frontline guardians against harmful content in the online world (Roberts, 2014; Bellanova & De Goede, 2021).

Despite the preference for using CMs as a means of self-regulation, a report from the Canadian Centre for Child Protection (CCCP) found:

> "[E]xpecting ESPs [Electronic Service Providers] to voluntarily invest the resources needed to reduce the availability of CSAM is simply not working … with unacceptably long delays for removing flagged content and previously flagged content re-emerging on websites, it is clear [they] are collectively failing to prioritise the safety and privacy of children online" (2021, p. 2).

The CCCP report noted that, while the median time for removal of content was 24 hours, around 10 percent remained available for seven weeks or longer before finally being removed, suggesting procedural deficiencies in the current system (CCCP, 2021, p. 4). Under Section 230 of the United States *Communications Decency Act* "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider"; essentially, Section 230 indemnifies technology companies, ISPs, and social media platforms from civil liability for the content posted by their users, as well as state-based criminal prosecution. There are some exceptions to this protection, for federal crimes like the proliferation of CSAM. Social media companies are criminally liable

for CSAM posted by users, but only if it can be proven that the company knew it was there and did nothing to remove it in a timely manner; this, supplemented by the overarching principles of CSR, drives both the content moderation process and, where appropriate, the referral of content to government agencies (Krishna, 2021).

On this, American social media companies are required by law to refer CSAM to the National Center for Missing and Exploited Children (NCMEC) or risk significant fines. The NCMEC works in conjunction with these companies, non-governmental organisations, the public, and law enforcement to coordinate responses to CSAM (among other things), and to support police action, both in the United States and abroad (Krishna, 2021; Bischoff, 2022). It is then the role of police to use referred content to (a) identify potential victims and/or offenders; (b) pass this intelligence to the authorities wherever those identified are located and; (c) coordinate to take legal action against offenders, based on legislation in the relevant territorial jurisdiction (Martellozzo, 2015; Bleakley, 2022).

Methodology

This article was designed to respond to two key questions related to the policing of CSAM on social media. First, the article examines current challenges presented by adapting traditional police methods (e.g. public law enforcement) to address the sharing of CSAM content online. The second overarching question explores the extent to which private social media companies have effectively become "the frontline" in dealing with CSAM via the process of content moderation. The purpose of this research is not to provide "solutions" to the challenges or problems identified; instead, the objective is to critically assess current academic and grey literature on the subject, in addition to other secondary sources (e.g. political statements; law

review articles etc), that can help provide better understanding of what the key issues are for both practitioners and researchers. For this reason, a scoping review was designed to meet the objective of the task more effectively through providing a critical "map of the evidence" (Munn et al, 2018, p. 143).

Initially, key terms were selected to guide initial searches, which were agreed among authors based on the research questions posed, and informed by subject matter expertise. Google Scholar was chosen as a search platform for its proven cross-disciplinary potential, which research suggests makes it a more appropriate tool for inter-disciplinary research than more subject-specific search engines (Lehnen & Insua, 2022). This proved effective, with relevant academic literature ultimately sourced from journals and other materials from a diverse range of subject areas including business, technology, public policy, and CSAM. Search terms were entered into Google Scholar, with the search filtered to only return results within a ten year period (2012-2022); this was considered essential to maintain relevance, particularly given the constantly changing nature of the Internet, social media, and online regulation. Specific search terms used were as follows: "content moderation"; "online harms"; "social media moderation"; online+CSAM, and; child+sexual+abuse+material.

This process yielded a large set of literature for critical analysis, requiring an extended exclusion process to identify the most useful literature for the purposes of this scoping review. Cross-referencing occurred with the purpose of eliminating repeated literature, which may have appeared in multiple searches. Due to the language capabilities of the research team, only English-language articles were included in this review. Non-relevant literature was also excluded at this point — non-relevance here is defined as literature not dealing with either: (a)

private social media content moderation; (b) policing and/or content moderation of CSAM on social media, and/or; (c) attempts to increase the regulation of social media platforms. Following this, the remaining literature was categorized based on its central focus or 'area of concern.' The result was that three foci – territoriality, privacy, and liability – emerged as the most prominent themes identified. Ancillary material was then eliminated which, while important to ensure a cohesive focus for this article, nevertheless also results in some limitation in providing a comprehensive, holisitic perspective on all issues related to the subject. We have attempted to address this using grey literature via a snowball method, as discussed below. Ultimately, the process described resulted in a shortlist of 36 relevant peer-reviewed journal articles, books, and chapters on the topic that were published between 2012 and 2022.

A secondary phase of the scoping process incorporated materials classified as 'grey literature' collected using a snowball method, which was informed by the authors' initial reading of the literature that was identified during the first phase of the scoping process (Jalali & Wohlin, 2012). This was seen as essential because of the emergent nature of the issues considered: like so many issues associated with the online world, the policing of CSAM on social media is an area constantly changing to a degree that academic research struggles to keep pace (Sarre, Lau & Chang, 2018). The grey literature used in this article derives from (and describes) laws, regulations, and policy debates from a European, British, Canadian, Australian, or American perspective. The reason for this is twofold. First, because our research team included only English-language products in our initial search, based on the linguistic capabilities of the research team. This same limitation meant that much of the grey literature that we gathered as part of the snowball process was also English-language, and focused on these contexts. On a more pragmatic level, drawing on grey literature from these contexts made sense given that these countries constitute some of the main geographic locations where social media companies

are headquartered and, as such, the legal and policy debates in these nations are often the most consequential for social media platforms in their provision of content moderation, and other online safety measures.

When discussing questions around the legalities of policing CSAM, changes in legislation or a private companies' terms of service can dramatically shift the landscape in a manner that outpaces the publication of academic research. To ensure the most current information was included in this review, non-peer reviewed material was considered, including: political statements; expert news articles and/or opinion pieces; law review articles; the text of legislation (proposed and enacted) and; technology company and social media platforms' policies, corporate documents, and user terms and conditions. These contextual sources were identified in a targeted manner, based on the results of the primary scoping review of the literature, through which the key issues were identified. This used a snowball approach: for example, when the issue of liability in relation to content moderation was noted in research by Roberts (2014), it prompted a more targeted exploration and resulted in recent moves to reform Section 230 in the United States being identified as an area of key debate, with relevant sources (e.g. proposed legislation, expert legal opinions) being incorporated into the final review. The contextualisation process was highly important to ensure that all material included was current at the time of writing.

Findings

Non-territoriality and the policing of harmful content in cyberspace

Conventional policing has always been closely tied to matters of geographical territory. Even when conceptualised in the broadest sense, legal jurisdiction is typically limited by physical location — where the crime took place, where the offender is located, or where the victim is located (Broadhurst, 2006; Perloff-Giles, 2018). With territorial jurisdiction underpinning the standard operating procedures of most global policing agencies, how well such a model can be transferred and applied in the online world remains unclear. Perloff-Giles (2018) notes that "when the perpetrator of such [cyber]crimes is located in the same jurisdiction as the victim, prosecution is relatively straightforward" (p. 200); for example, when CSAM material is shared via online channels, police are often (albeit not easily) able to use contextual clues embedded in the image and/or video to identify where the content was created and, from there, initiate police action (Bleakley, 2022).

Once territorial jurisdiction is established, 'normal' law enforcement can resume: contact offenders typically reside in the same geographical area as their victims, which is also usually where the crime itself took place (Cale et al., 2021). To establish and act on territorial jurisdiction over offences depicted in online CSAM content is a more straightforward process than combatting the dissemination of that content. When CSAM is posted online, one of the first steps from law enforcement is to determine the sharer's geographic location — a task that is obstructed by considerable barriers, such as the inherent anonymity of the online world, and the easy availability of technology designed to hide users' true physical location (e.g. VPNs, Tor browsers) (Henderson, 2020). Even where a user's identity can be ascertained, the trans-national nature of social media means there is a relatively slim likelihood that the person is

physically located within their legal jurisdiction. Questions are then raised as to which jurisdiction a person *is* subject to when transmitting CSAM over social media: (1) that of their physical location; (2) the location where the material is published (or downloaded), or; (3) that of the publisher — in this context, the jurisdiction where social media platforms are based?

Each of these is problematic. The simplest, perhaps, is to consider an offender subject to the jurisdiction where they committed the offence, or where they are physically located. The position of the Crown Prosecution Service (CPS) in England and Wales is that "an offence will only be triable in the jurisdiction in which the offence takes place", except for limited contexts where legislation permits the application of extra-territorial jurisdiction (Crown Prosecution Service [CPS], 2018). This position is common among global legislative frameworks and is, perhaps, a recognition of the practicalities of investigating and prosecuting trans-national, online crime. Ensuring that individuals sharing illegal content are dealt with effectively in their own territorial jurisdiction requires police to work closely with international law enforcement partners in sharing intelligence, and collaborate on multinational investigations (Holt et al., 2020). Where laws on illegal content (like CSAM) are aligned, as the European Union *Digital Services Act* aims to achieve among member states, this international cooperation can be effective; however, the nuances of what precisely constitutes illegal content in various countries around the world, as well as varied statutory restrictions on police investigatory practices, can complicate interagency cooperation (Bleakley, 2019; Holt et al., 2020). The second method, to determine jurisdiction based on where material is published, is far more problematic. On this issue, Perloff-Giles quotes Brierly's (pre-Internet) view that "the suggestion that every individual is or may be subject to the laws of every State at all times and in all places is intolerable" (2018, p. 205). While individuals are certainly subject to the laws of a foreign country where they physically travel and commit a crime, an individual who posts

content on social media "is not consciously choosing to bind themselves to any particular foreign government's laws" in the same way (Daskal, 2015, pp. 367-368). Beyond this existential argument, there are also significant practical challenges involved in holding individuals accountable for crimes when they are not physically present in the jurisdiction where they face charges, making this approach generally unsustainable from a policing standpoint.

This leaves the third option — to establish jurisdiction as the location where the publishing entity, the social media platform, is physically located. With so many of these corporate entities based in the United States, where Section 230 of the *Communications Decency Act* shields social media companies from much liability over what is posted on their sites, the legal framework to support this approach is fundamentally weakened (Salter & Hanson, 2021). There is, however, some precedent for adopting an approach treating jurisdiction as equivalent to the physical location of the technology infrastructure that 'hosts' illegal materials. A controversial element of the Queensland Police's investigation into Dark Web CSAM site *Child's Play* involved the agency copying and transferring the entire website from its original server location to a different server in Australia, allowing them to operate a months-long sting operation on the site in a jurisdiction that permitted it (Hoydal, Stangvik and Hansen, 2017; Bleakley, 2019). Although this move attracted criticism of police "jurisdiction shopping", it remains a prime example of law enforcement basing legal jurisdiction on server location. While this option may streamline some of the law enforcement issues linked to offending on social media, jurisdictional complications remain, such as the inability for police to take action against individuals in other territories; as such, a truly viable approach to jurisdiction on social media (and in cyberspace, more generally) continues to be elusive.

Privacy, free speech, and the neutrality of a liberal 'cyberspace'

In the digital realm CMs have taken on an essential role as the chief arbiters on UGC, tasked with determining whether said material is appropriate, lawful, and should be hosted on their employer's social media platform. In part, this is due to the aforementioned challenges around jurisdiction: whereas online jurisdiction can often prove complicated to establish for law enforcement, a social media platform is the only entity that truly has full "jurisdiction" in regard to what is posted on its platforms (Edwards et al., 2021). In a practical sense, this makes CMs the most logical group to serve as a first line of defence when it comes to CSAM, with direct power to intervene with immediacy as soon as problematic material is identified. Taking on this role also allows social media companies to showcase their commitment to CSR, and ensure a positive experience for users — which, in turn, grows and retains a robust customer-base that is appealing to advertisers, and ensures a continued revenue stream for online service providers (Myers West, 2018). Critics argue that the process of corporate-led content moderation implicitly contradicts principles of free speech online, with technology companies like Meta or Twitter "not a governmental actor … reliev[ing them] of all formal constitutional concerns about [their] content restriction policies" and raising concerns around "how much discretion should be allowed to the censor" (Langvardt, 2017, p. 1356).

Discussing the enforcement of social media content moderation rules, Common (2020) observes that "the current approach adopted by most platforms is underdeveloped … [and reflects a] bias in decision-making, an over-reliance on efficiency as a solution, and inconsistent enforcement of terms and conditions" (p. 126). Langvardt (2017) believes this "danger to free expression is amplified today because contemporary Internet platforms

comprehend and mediate a far larger share of communications than was previously possible … [online platforms] must move aggressively and quickly to suppress an enormous volume of unwanted communications" (p. 1360). The sheer volume of material CMs are required to respond to requires policies and protocols favouring speed and efficiency in censorship — argued by some as contributing to an over-zealous response that crosses a line, ultimately infringing on general freedom of speech (Langvardt, 2017; Land, 2019). Concerns over CMs perceived role as unregulated online censors are such that several U.S. states have attempted to introduce laws "that would prevent digital services from moderating harmful content to protect their users" by severely curtailing the types of content that technology companies are allowed to moderate, on free speech grounds (Greenfield, 2022). As the legal status quo currently stands, content moderation is protected with social media companies not beholden to the same strict legal requirements to allow unfettered expression as the state would be; again, this is another argument in favour of self-regulation, with the law affording private organisations greater power to police online content in many respects than it does traditional law enforcement (Edwards et al., 2021).

This debate around CMs' censor role has focused on disinformation, political, and/or hate speech, based on the argument that these communications represent protected (or, at least, quasi-protected) expression. The same argument does not extend to other forms of harmful content, such as CSAM, which do not constitute protected speech and are subject to existing federal laws requiring private companies to remove content in a timely manner or risk criminal penalties (Krishna, 2021). Even so, prevention of CSAM has been cited as a motivating factor in another debate over privacy and free speech online: the campaign against end-to-end encryption (E2EE). The purpose of E2EE is to provide enhanced privacy for the clients of technology companies, assuring them that their communications are secured and nobody – not

even law enforcement, or the platform itself – is able to surreptitiously 'spy' on their conversations (Watney, 2020). International efforts to strengthen anti-CSAM practices, including the United States EARN IT Act (first introduced in 2020), argue that E2EE provides bad actors with an enhanced ability to share CSAM via digital platforms without detection and, as such, mechanisms need to be built in to allow law enforcement to access and/or scan communications, breaching encryption (Eoyang & Garcia, 2020).

Opponents to these reforms claim allowing CMs (or law enforcement) to scan E2EE communications for illegal content like CSAM is a slippery slope: as Kamara et al. (2021) correctly points out, "the types of content of interest – typically 'harmful,' illegal, or otherwise unwanted content such as terrorist propaganda, CSAM, mis- and disinformation, or spam – have no technically unique characteristics … thus, what is often framed as a debate about moderation of unwanted content in E2EE services is really a discussion about (any) content detection in E2EE" (p. 16); in short, to allow scanning of E2EE for illegal or harmful content would be tantamount to ending the protections of E2EE entirely, severely compromising user privacy. Once again, proponents of E2EE argue in favour of traditional self-regulatory preferences, asserting that user reporting and metadata analysis are the best ways to respond to the potential for illegal content to be shared via E2EE, without having to rely on CMs to actively scan correspondence and breach users' online privacy (Watney, 2020; Kamara et al., 2021). These 'solutions' are not without their own problems: user reporting would not, for example, resolve the sharing of CSAM between two complicit parties through E2EE, while it is possible to circumvent the machine learning involved in metadata analysis, with appropriate technical skill. Ultimately, current best practice dictates that there is an essential need for the 'human factor' in content moderation to make decisions on content in line with a platform's

community standards and, for now, this will continue to raise serious concerns around privacy and censorship that are difficult to resolve on a philosophical level.

Social media, Section 230 and the 'Liability Dilemma'

An often-overlooked dimension to regulating online communications (or, "free speech") is the fact that, as Ardia (2009) puts it, "what many consider the largest public space in human history [the Internet] is not public at all … it is layered on privately owned Web sites, privately owned servers, privately owned routers, and privately owned backbones [and] without the acquiescence of these intermediaries, the public would have no access to speak or to be heard" (p. 377). Because of this, the Internet is a space where service providers are largely free from the First Amendment restrictions that would otherwise limit their legal ability to moderate content posted on their platforms; on the flipside, however, the same companies would nevertheless be *liable* for harmful content published on their site if not for the provisions of Section 230(c)(1) of the United States *Communications Decency Act*.[1] Introduced in 1996, this section was designed as a form of 'Good Samaritan' protection that would shield an online operator from civil and (some) criminal liability for the actions of others, in this case the users of their platform and/or service (Sevanian, 2014). Typically, a publisher would be held liable for harmful or illegal content (e.g., in print media); under Section 230(c)(1), "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." The section has been credited widely as an essential step in the development of the Internet, relieving online entrepreneurs from assuming burdensome risks of legal action. Since its introduction in 1996, other

[1] The provisions of the *Communications Decency Act* apply only within the territorial jurisdiction of the United States – however, as most of the major social media providers are headquartered in the U.S., the legislation has great influence over global operations in the sector.

international jurisdictions have adopted similar legislation, as in the "safe harbor" provisions of the European Union Directive 2000/31/EC (Bessen & Verveer, 2021).

For most of the Internet age, Section 230 has been consistently upheld whenever online platforms have been faced with legal challenge over UGC (*Doe v. America Online*, 2001; Marin & Popov, 2007). Though social media platforms are indemnified from civil claims over UGC, as well as same state claims, Section 230 does not provide protection for certain federal crimes, in particular knowingly-hosting CSAM. Under federal law, a social media platform is *not* liable for CSAM being posted, but is liable if it can be shown that they were aware it was on their site and did not take reasonable steps to remove and report it to authorities (Ardia, 2009; Krishna, 2021). While this is important in that it limits Section 230 protections, this status quo is still predicated on a self-regulation wherein operators are encouraged to be reactive to content, rather than proactive in preventing it from appearing in the first place. With Canada-based Project Arachnid finding a small, yet significant, proportion of delays in removal of CSAM (as well as exponentially growing volume of illicit content being shared on social media), critics of Section 230 suggest that social media providers are not adequately incentivised to act swiftly on harmful content like CSAM, partly as a result of their quasi-protected legal status (CCCP, 2021; Meta, 2022). These criticisms drove reforms in 2018 to limit Section 230 for online providers found to be knowingly (or unknowingly) facilitating sex trafficking under FOSTA-SESTA laws (Mikaelyan, 2021).

As Ardia (2009) observed, Section 230 is dichotomously characterised by proponents "as the saviour of free speech in the digital age and [by opponents] as an ill-conceived shield for scoundrels" (pp. 379-380). In his first State of the Union address in March 2022, United States

President Joe Biden criticised social media companies for running a "national experiment" on young users, and pledged to hold them accountable — though he did not name Section 230 reforms specifically, any attempts to increase these platforms' liability would inevitably require some watering-down of existing protections under the act (Ghaffary, 2022). Obama-era Commerce official Cameron F. Kerry acknowledged (potential) unforeseen consequences from well-intentioned attempts to hold social media companies legally liable: he argues that a legal mandate to moderate content – which is already being done on a self-regulatory basis – may result in CMs being forced to make the difficult choice "between having to allow any kind of offensive content onto their platforms or facing liability for offensive content that slips through" (Kerry, 2021).

Due to strict legal liability being applied with a reform or repeal to Section 230, major social media providers like Meta and Twitter would become more reliant on automated, algorithm-based technologies to ensure that no illicit content is posted that would expose them to legal risk; reliance on such technology in itself presents further risks of over-censorship, with machine-learning unable to provide the necessary "contextual judgements [made] by humans to identify the problems and decide what's over the line" (Kerry, 2021). However, use of artificial intelligence (AI) to mechanize the process of social media moderation is not without its own legal issues. As Panait and Ashraf (2021) note, the European General Data Protection Regulation (GDPR) includes provisions against reliance on AI which state that an individual cannot be subject to decisions based solely on automated processes. This regulation is part of broader GDPR protections around privacy and, in part, was based on fears that automated moderation could prove overly censorious, and restrictive in regard to private communications online (Bharti & Aryal, 2022). While affirming the importance of the human CM in this process, even with AI support, Panait and Ashraf acknowledge that the GDPR has limited

regional impact, and has not proven effective in disincentivizing an increasingly expansive use of AI as a tool in moderation by social media platforms.

Additional legal analysis suggests that self-regulation is the only approach to content moderation that is effective from a constitutional standpoint, under U.S. law: because social media providers currently undertake content moderation on a voluntary basis, all material CMs detect and refer to authorities is considered admissible in court. If the law is changed to *compel* social media providers to actively search for illegal material on their platforms, this all-important voluntarism (or, self-regulation) aspect would cease to apply, and evidence gathered by social media CMs – now under legal compulsion to act – may be seen as an unlawful search and seizure under the United States Fourth Amendment, potentially meaning evidence that is gathered will be inadmissible (Zabel, 2020; Nuthi, 2022). It must be noted that these legal complications are currently hypothetical — without being tested in practice, it is uncertain what true impact reform and/or repeal of Section 230 would have on content moderation.

Conclusions and areas for further consideration

When the Internet was still in its infancy, self-regulation by online service providers like social media platforms was seen as a practical solution to the challenges posed by the new digital landscape. Without clarity on basic questions like how to manage competing, overlapping, or otherwise unclear jurisdiction – as well as an overarching philosophical preference for an uber-liberal "cyberanarchy" (Weiser, 2001) – permitting cyberspace to operate with minimal interference was not just an appealing option but, in many ways, the *only* viable option available at the time. However, as the role of the Internet has exponentially expanded to the point that it has become such an intrinsic element of society, this self-regulation has had a

substantial impact on public safety. Where before the state and its law enforcement apparatus was responsible for performing this role, the balance has tipped in the Internet age to the point that social media platforms like Facebook, Twitter and TikTok now serve as the first line of defence against harmful content like CSAM, with CMs serving as first responders for the digital age (Roberts, 2014; Bellanova & De Goede, 2021).

This paradigm shift brings with it new challenges and, as such, necessitates a greater interrogation of whether existing legislation, systems, and practices are properly equipped to respond to the proliferation of CSAM. This includes consideration of CMs ability to cope with the negative social, emotional, and psychological repercussions of the profession: with AI not having developed to the point that renders human moderation redundant, ensuring that CMs are adequately equipped to perform their duties in a way that prioritizes their own personal well-being remains an area requiring further research, and greater intervention on an industry level (Spence et al., 2023). Underpinning all is the existential question of whether "outsourcing" policing of the online world to social media companies remains the most effective mode of regulation and, if so, if modifications to processes and protocols need to be made to make existing systems more efficient.

The recent European agreement on enhanced online regulation may be a critical first step in reforming, or even dismantling, the existing state of digital self-governance. Agreed to in principle in April 2022, the *Digital Services Act (DSA)* was created on the basis that "what is illegal offline must also be illegal online" (European Parliament, 2022b) — while the law's final text has yet to be confirmed and approved, the *DSA* will establish a "notice and action mechanism" requiring online services to act "without undue delay [on reports of harmful

content], taking into account the type of illegal content that is being notified and the urgency of taking action" (European Parliament, 2022a). Failure to comply with provisions like the timely removal of CSAM may result in financial penalties for online platforms, as high as 6 percent of worldwide fiscal turnover (Beer, 2022; European Parliament, 2022b). The *DSA* came into force on 16 November 2022, however affected service providers have until 1 January 2024 to comply with the new regulations (European Parliament, 2022c). Even then, it will be difficult to assess its impact on content moderation process with veracity for some time. Critics argue the *DSA* is destined to fail "without effective and properly functioning enforcement … [and] remain an empty shell" (Pirkova, 2021). Even with the ability to enact stronger punitive measures against non-compliant digital platforms, *DSA* skeptics argue that it will only function effectively as part of a cooperative model, relying on technology companies continuing to serve as the proactive frontline of content moderation — essentially, a replication of structures and processes in the current status quo.

One of the key reasons that social media platforms, and their CMs, continue to assume such a central position in policing cyberspace emerges from somewhat outdated notions of territorial jurisdiction. The global interconnectivity of the Internet confounds these traditional hallmarks of policing, resulting in social media platforms the only entity with total "jurisdiction" over UGC (Roberts, 2014; Myers West, 2018; Bellanova & De Goede, 2021). While content moderation by those same platforms may, thus, be the most pragmatic response, it nevertheless raises concerns around the abdication of the state's traditional authority to regulate the activities of its citizens to private technology companies. Critics contend that such responsibility (and authority) must come with greater regulation — for example, laws to limit E2EE as a means of better combatting CSAM being sent via secure social media communications. Amongst those entering this policy debate is the former American President Barack Obama who, in his latest

speech at Stanford's Cyber Policy Center, highlighted the importance of content moderation, arguing that unregulated social media presents a threat to democracy. He urges people to "pick a side" on whether social media companies should be regulated by the government and made more responsible for the content published on their services (Pearce, 2022).

These attempts to regulate face their own criticism, however: arguments that such actions would fundamentally breach users' privacy online have frustrated reform campaigns (Kamara et al., 2021). These legitimate concerns derive largely from the conflation of overtly harmful content like CSAM with other forms of "speech" such as political discourse and misinformation — whereas a majority may be in favour of government intervention designed to limit the dissemination of CSAM, when the measures necessary to do so would also impact on the limitation of "free speech" in other areas (as is the case with the debate on E2EE), it frustrates these efforts to prevent CSAM, effectively causing it to become collateral damage in a larger fight over the control of online speech in a more general, and contentious, sense.

Other reforms centred on increasing technology companies' liability for UGC (through amendment to, or repeals of, Section 230 or similar legislation) have also been met with pushback from opponents, who argue that well-intentioned efforts to strengthen responses to CSAM would only disrupt the already-successful process of self-regulated content moderation, potentially resulting in CMs becoming over-sensitive to questionable content in a way that may repress freedom of expression (Kerry, 2021). While attempts at greater governance like the *DSA* have attained widespread support, and seen a degree of legislative success, there remain questions as to whether it is inevitable that a lack of jurisidction (and, thus, control) over social media platforms exerted by any one body will frustrate efforts to enforce compliance. If social media platforms are to remain the frontline against harmful content, it must first be

acknowledged that this represents a seismic shift in the traditional policing model; once this is acknowledged, we may be able to better address the myriad of issues that this "new normal" raises and, from there, work to improve systems and processes in a way that balances the need for securitisation with the inherent rights of the public.

References

Ardia, D. S. (2009). Free speech savior or shield for scoundrels: An empirical study of intermediary immunity under Section 230 of the Communications Decency Act. *Loyola of Los Angeles Law Review, 43*, 373-506.

Banko, M., MacKeen, B., & Ray, L. (2020). A Unified Typology of Harmful Content. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 125-137. http://dx.doi.org/10.18653/v1/2020.alw-1.16.

Beer, E. (2022). The EU's Digital Services Act — what you need to know. *The Stack*, 26 April. Available from https://thestack.technology/eu-digital-services-act-what-you-need-to-know/ (accessed 27 April 2022).

Bellanova, R., & De Goede, M. (2021). Co-Producing Security: Platform Content Moderation and European Security Integration. *Journal of Common Market Studies*, 1-19. https://doi.org/10.1111/jcms.13306.

Bessen, S. M., & Verveer, P. L. (2021). Section 230 and the problem of social cost. *Journal of Law and Policy, 30*(1), 68-120.

Bharti, S. S., & Aryal, S. K. (2022). The right to privacy and an implication of the EU General Data Protection Regulation (GDPR) in Europe: challenges to the companies. *Journal of Contemporary European Studies*, https://doi.org/10.1080/14782804.2022.2130193.

Bischoff, P. (2022). The rising tide of child abuse content on social media. *Comparitech,* 11 January. Available from https://www.comparitech.com/blog/vpn-privacy/child-abuse-online-statistics/ (accessed 10 March 2022).

Bleakley, P. (2019). Watching the watchers: Taskforce Argos and the evidentiary issues involved with infiltrating Dark Web child exploitation networks. *The Police Journal: Theory, Practice and Principles, 92*(3), 221-236.

Bleakley, P. (2022). *Policing Child Sexual Abuse: Failure, Corruption and Reform in Queensland.* London, UK: Routledge.

Brown, I. (2010). Beware self-regulation. *Index on Censorship, 39*(1), 98-106.

Bursztein, E., Clarke, E., DeLaune, M., Elifff, D. M., Hsu, N., Olson, L., & Bright, T. (2019, May 13–17). Rethinking the detection of child sexual abuse imagery on the internet. *Proceedings of the 2019 world wide web conference*, WWW '19, San Francisco, CA.

Cale, J., Holt, T., Leclerc, B., Singh, S., & Drew, J. (2021). Crime commission processes in child sexual abuse material production and distribution: a systematic review. *Trends and Issues in Crime and Criminal Justice,* (617), 1-22.

Canadian Centre for Child Protection [CCCP]. (2021). *Project Arachnid: Online Availability of Child Sexual Abuse Material: Summary Document*. Winnipeg, Manitoba: Canadian Centre for Child Protection.

Clement, J. (2019). Number of social network users worldwide from 2010 to 2021 (in billions). *Statista*. Available from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users (accessed 8 March 2022).

Communications Decency Act of 1996, 47 U.S.C. § 230.

Council of Europe. (1993). *Sexual exploitation, pornography and prostitution of, and trafficking in, children and young adults: Recommendation No. R(91) 11 and Report of the European Committee on Crime Problems.* Strasbourg: Council of Europe.

Crown Prosecution Service [CPS]. (2018). Social media – Guidelines on prosecuting cases involving communications sent via social media. *Crown Prosecution Service.* Available from https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media (accessed 10 March 2022).

Cusumano, M. A., Gawer, A., & Yoffie, D. B. (2021). Can self-regulation save digital platforms? *Industrial and Corporate Change, 30*(5), 1259-1285.

Dabrowska, M. (2021). The unclear picture of child sexual abuse material (CSAM) online volumes during the COVID-19 pandemic. *Bialystok Legal Studies, 26*(6), 109-125.

Daskal, J. (2015). The Un-Territoriality of Data. *Yale Law Journal*, *125*, 326-398.

Drootin, A. (2021). "Community Guidelines": The legal implications of workplace conditions for Internet content moderators. *Fordham Law Review, 90*(3), 1197-1244.

Edwards, A., Webb, H., Housley, W., Beneito-Montagut, R., Procter, R., & Jirotka, M. (2021). Forecasting the governance of harmful social media communications: findings for the digital wildfire policy Delphi. *Policing and Society, 31*(1), 1-19.

Eoyang, M., & Garcia, M. (2020). *Weakened encryption: The threat to America's national security.* Washington, D.C.: Third Way.

European Parliament (2022). Digital Services Act: regulating platforms for a safer online space for users. *European Parliament,* 20 January. Available from https://www.europarl.europa.eu/news/en/press-room/20220114IPR21017/digital-services-act-regulating-platforms-for-a-safer-online-space-for-users (accessed 27 April 2022).

European Parliament. (2022). Digital Services Act: agreement for a transparent and safe online environment. *European Parliament*, 23 April. Available from https://www.europarl.europa.eu/news/en/press-room/20220412IPR27111/digital-services-act-agreement-for-a-transparent-and-safe-online-environment (accessed 27 April 2022).

European Parliament. (2022). Digital Services: Landmark rules adopted for a safer, open online environment. *European Parliament*, 5 July. Available from https://www.europarl.europa.eu/news/en/press-room/20220701IPR34364/digital-services-landmark-rules-adopted-for-a-safer-open-online-environment (accessed 20 March 2023).

Ghaffary, S. (2022). Biden threatens Big Tech over its "national experiment" on children. *Vox*, 1 March. Available from https://www.vox.com/recode/2022/3/1/22957507/biden-state-of-the-union-social-media-mental-health-children-accountability-frances-haugen (accessed 14 March 2022).

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society, 7*(2), 1-5.

Greenfield, H. (2022). Georgia's Senate passes content moderation bill similar to those courts have found unconstitutional. *Computer & Communications Industry Association*, 9 March. Available from https://www.ccianet.org/2022/03/georgias-senate-passes-content-moderation-bill-similar-to-those-courts-have-found-unconstitutional/ (accessed 11 March 2022).

Gunningham, N., & Rees, J. (1997). Industry self-regulation: An institutional perspective. *Law & Policy, 19*(4), 363-414.

Henderson, L. (2020). *Tor and the Dark Art of Anonymity*. Seattle, WA: Amazon.

Holt, T. J., Cale, J., Leclerc, B., & Drew, J. (2020). Assessing the challenges affecting the investigative methods to combat online child sexual exploitation material offenses. *Aggression and Violent Behavior, 55*, 1-7. https://doi.org/10.1016/j.avb.2020.101464.

Hoydal, H. F., Stangvik, E. O., & Hansen, N. R. (2017) Breaking the Dark Net: Why the police share abuse pics to save children. *VG*, 17 October. Available from http://www.vg.no/spesial/2017/undercover-darkweb/? (accessed 17 February 2022).

Jalali, S., & Wohlin, C. (2012). Systematic literature studies: Database searches vs. backward snowballing. *Proceedings of the 2012 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 2012, 29-38. doi:10.1145/2372251.2372257.

Johansson, C. (2019). *Combating online child sexual abuse material: An explorative study of Swedish police investigations* (Master's thesis). Malmo University, Malmo, Sweden.

Kamara, S., Knodel, M., Llanso, E., Nojeim, G., Qin, L., Thakur, D., & Vogus, C. (2021). *Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems*. Washington, D.C. & Brussels, Belgium: Center for Democracy and Technology.

Kerry, C. F. (2021). Section 230 reform deserves careful and focused consideration. *Brookings Institute*, 14 May. Available from https://www.brookings.edu/blog/techtank/2021/05/14/section-230-reform-deserves-careful-and-focused-consideration/ (accessed 14 March 2022).

Krishna, A. (2021). Internet.gov: Tech companies as government agents and the future of the fight against child sexual abuse. *California Law Review, 109*(4), 1581.

Land, M. K. (2019). Against privatised censorship: Proposals for responsible delegation. *Virginia Journal of International Law, 60*(2), 363-432.

Langvardt, K. (2018). Regulating online content moderation. *The Georgetown Law Journal, 106*(5), 1353-1388.

Lehnen, C. A., & Insua, G. M. (2022). Search tools and scholarly citation practices in literary studies. *Reference Services Review, 0*(0), https://doi.org/10.1108/RSR-07-2022-0025.

Mac Síthigh, D. (2017). *Medium Law*. Abingdon, Oxon: Routledge.

Marin, M. D., & Popov, C. V. (2007). Doe v. MySpace, Inc.: Liability for third party content on social networking sites. *Communications Lawyer, 25*, 3.

Martellozzo, E. (2015). Policing online child sexual abuse – the British experience. *European Journal of Policing Studies, 3*(1), 32-52.

McIntyre, T. J. (2013). Child abuse images and cleanfeeds: Assessing Internet blocking systems. In I. Brown (Ed.). *Research Handbook on Governance of the Internet* (pp. 277-308). Cheltenham, UK: Edward Elgar.

Meta. (2022). Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation [Facebook]. *Meta Transparency Center*. Available from https://transparency.fb.com/data/community-standards-enforcement/child-nudity-and-sexual-exploitation/facebook/ (accessed 8 March 2022).

Mikaelyan, Y. (2021). Reimagining content moderation: Section 230 and the path to industry-government cooperation. *Loyola of Los Angeles Entertainment Law Review, 41*(2), 179-214.

Munn, Z., Peters, M. D. J., Stern, C., Tufarnaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*(1), 143.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society, 20*(11), 4366-4383.

Nolte, A., & Westermeier, C. (2020). Between Public and Private: The co-production of infrastructural security. *Politikon, 47*(1), 62-80.

Nuthi, K. (2022). The EARN IT Act would give criminal defendants a get-out-of-jail-free card. *Slate*, 11 February. Available from https://slate.com/technology/2022/02/earn-it-act-fourth-amendment-violation.html (accessed 1 March 2022).

Obar, J. A., & Wildman, S. S. (2015). Social media definition and the governance challenge – an introduction to the special issue. *Telecommunications Policy*, *39*(9), 745-750.

Panait, C., & Ashraf, C. (2021). AI algorithms – (re)shaping public opinions through interfering with access to information in the online environment? *Europuls Policy Journal, 1*(1), 46-64.

Pearce, M. (2022) Obama argues unregulated social media is a threat to democracy, calls to 'pick a side.' *Los Angeles Times*, 21 April. Available from https://www.latimes.com/entertainment-arts/story/2022-04-21/la-ent-obama-disinformation-stanford (accessed 22 April 2022).

Perloff-Giles, A. (2018). Transnational Cyber Offenses: Overcoming Jurisdictional Challenges. *Yale Journal of International Law, 43*(1), 191-228.

Pirkova, E. (2021). The EU Digital Services Act won't work without strong enforcement. *Access Now*, 9 December. Available from https://www.accessnow.org/eu-dsa-enforcement/ (accessed 27 April 2022).

Roberts, S. T. (2016). Commercial Content Moderators: Digital Laborers' Dirty Work. In S. U. Noble, & B. M. Tynes (Eds.). *The Intersectional Internet: Race, Sex, Class, and Culture Online* (pp. 147-160). New York, NY: Peter Lang.

Roberts, S. T. (2016). Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste. *Wi: Journal of Mobile Media, 10*(1), 1-11.

Salter, M. (2017). *Crime, Justice and Social Media.* London; New York, NY: Routledge.

Salter, M., & Hanson, E. (2021). "I need you to understand how pervasive this issue is": User efforts to regular child sexual offending on social media. In J. Bailey, A. Flynn, & N. Henry (Eds.). *The Emerald International Handbook of Technology-Facilitated Violence and Abuse* (pp. 729-748). Bingley, UK: Emerald.

Sarre, R., Lau, L. Y., & Chang, L. Y. C. (2018). Responding to cybercrime: current trends. *Police Practice and Research, 19*(6), 515-518.

Sevanian, A. M. (2014). Section 230 of the Communications Decency Act: A "Good Samaritan" law without the requirement of acting as a "Good Samaritan." *UCLA Entertainment Law Review, 21*(1), 121-146.

Spence, R., Harrison, A., Bradbury, P., Bleakley, P., Martellozzo, E., & DeMarco, J. (2023). Content moderators' strategies for coping with the stress of moderating content online. *Journal of Online Trust and Safety,* forthcoming.

Steiger, M. (2020). *Building a resilient workforce: Programming for Commercial Content Moderation Staff* (Doctoral thesis). St. Mary's University, San Antonio, TX.

Steiger, M., Bharucha, T.J., Venkatagiri, S., Riedl, M.J., & Lease, M. (2021). The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In CHI Conference on Human Factors in

Computing Systems (CHI ’21), May 8-13, 2021, Yokohama, Japan. DOI: 10.1145/3411764.3445092

Watney, M. (2020). Law enforcement access to end-to-end encrypted social media communications. In C. Karpasitis & C. Varda (Eds.). *Proceedings of the 7th European Conference on Social Media* (pp. 322-329). Reading, UK: ACPI.

Zabel, J. (2020). Public surveillance through private eyes: The case of the EARN IT Act and the Fourth Amendment. *University of Illinois Law Review*, *2020*(Fall), 167-177.