International Journal of Cybersecurity Intelligence & Cybercrime

Volume 1 | Issue 1 Article 3

8-2018

A Failure to Regulate? The Demands and Dilemmas of Tackling Illegal Content and Behaviour on Social Media

M. Yar. (2018). *International Journal of Cybersecurity Intelligence and Cybercrime, 1* (1), 5-20.

A Failure to Regulate? The Demands and Dilemmas of Tackling Illegal Content and Behavior on Social Media

Majid Yar*, Lancaster University, U.K.

Key Words: social media, cyber-bullying, cyber-terrorism, extremism, regulation, governance

Abstract:
The proliferation and user uptake of social media applications has brought in its wake a growing problem of illegal and harmful interactions and content online. In the UK context, concern has focused in particular upon (a) sexually-oriented content about or directed to children, and (b) content that is racially or religiously hateful, incites violence, and promotes or celebrates terrorist violence. Legal innovation has sought to make specific provision for such online offences, and offenders have been subject to prosecution in some widely-publicized cases. Nevertheless, as a whole, the business of regulating (identifying, blocking, removing, and reporting) offending content has been left largely to social media providers themselves. This has been sustained by concerns both practical (the amount of public resource that would be required to police social media) and political (concerns about excessive state surveillance and curtailment of free speech in liberal democracies). However, growing evidence about providers' unwillingness and/or inability to effectively stem the flow of illegal and harmful content has created a crisis for the existing self-regulatory model. Consequently, we now see a range of proposals that would take a much more coercive and punitive stance toward media platforms, so as to compel them into taking more concerted action. Taking the UK as a primary focus, these proposals are considered, with a view to charting possible future configurations for tackling illegal social media content.

Introduction

The proliferation and user uptake of social media applications has brought in its wake a growing problem of illegal and harmful interactions and content online. Recent controversy has arisen around issues ranging from the alleged online manipulation of the 2016 US presidential election by Russian hackers and "trolls," to the misuse of users' Facebook data by the political consulting firm Cambridge Analytica (Hall, 2018; Swaine & Bennetts, 2018). These recent issues notwithstanding, in the UK context, ongoing concern has focused in particular upon (a) sexually-oriented and abusive content about or directed at children, and (b) content that is racially or religiously hateful, incites violence and promotes

*Corresponding author
Majid Yar, Ph.D., Professor of Criminology, Law School, Lancaster University, Bowland North, Lancaster, United Kingdom, LA1 4YN.
Email: m.yar2@lancaster.ac.uk

or celebrates terrorist violence. Legal innovation has sought to make specific provision for such online offences, and offenders have been subject to prosecution in some widely-publicized cases. Nevertheless, as a whole, the business of regulating (identifying, blocking, removing and reporting) offending content has been left largely to social media providers themselves. This has been sustained by concerns both practical (the amount of public resources that would be required to police social media) and political (concerns about excessive state surveillance and curtailment of free speech in liberal democracies). However, growing evidence about providers' unwillingness and/or inability to effectively stem the flow of illegal and harmful content has created a crisis for the existing self-regulatory model. Consequently, we now see a range of proposals that would take a much more coercive and punitive stance toward media platforms, so as to compel them into taking more concerted action. Taking the UK as a primary focus, these proposals are considered and assessed, with a view to charting possible future configurations for tackling illegal social media content. However, this paper is not intended as an exercise in criminal justice policy analysis in the conventional sense, as in one that seeks to evaluate the effectiveness of a given policy and on that basis offer recommendations for policy reform or innovation. Rather, by taking a broader, socially-embedded criminological approach, it aims to elaborate an important emerging shift in the political and public discourse on online policing and crime control, and to explain this change by reference to the supposed limitations and/or failures of existing self-regulatory approaches to social media content.

Crime and Social Media: The Rise of an Online Problem

The early explosion in internet use (roughly dated to the early-1990s) was initially based upon a reproduction of communication structures and practices associated with established broadcast and print-based media channels (e.g., television, radio, newspapers, magazines and books). With the exception of email (enabling user-to-user communication) and discussion lists and bulletin boards (enabling collective sharing of text-based content), the vast majority of internet use entailed a one-way flow of communication from producers to consumers, i.e. internet use was primarily configured around "passiv" consumption of static content posted on websites without any meaningful input or interaction on the part of the content's readers or viewers. This era, which marks the first generation in the evolution of the Internet, is now referred to as "Web 1.0"' (Cormode & Krishnamurthy, 2008). In contrast, the late 1990s and early 2000s saw the emergence of the second iteration in the evolution of the web, so-called Web 2.0. What distinguished Web 2.0 platforms, applications, and services from their predecessors was their erosion (and possible effacement) of the distinction between producers and consumers of content, with users generating their own multi-media content (e.g., text, images, animations, audio-visual recordings), which could be shared with multiple other users. In this communicative configuration, traditional media consumers were replaced by "prosumers" (Ritzer, Dean & Jurgenson, 2012) or "producers" (Bird, 2011), individuals and groups who engaged in two-way, many-to-many interaction via a variety of online platforms. It is precisely this user interactivity that puts the "social" into "social media." Additionally, Web 2.0 was characterised by its accessibility and ease-of-use for those with little or no expertise in the technical aspects of computer use, as well as "interoperability" enabling applications to work seamlessly across different systems and devices (O'Reilly, 2010). Taken together, these features facilitated rapid growth and user uptake of Web 2.0, notable in the conspicuous popularity of the likes of MySpace (launched in 2003), Facebook (launched in 2004), and YouTube and Bebo (both launched in 2005) (Boyd & Ellison, 2007). While some of these initial success stories have largely fallen by the wayside in what is a dynamic and competitive market (MySpace, Friendster, and Digg being notable instances), overall the sector continues to grow and thrive, with the likes of Twitter, Flickr, Tumblr, LinkedIn, and WhatsApp each gathering hundreds of millions of users.

The aforementioned communication structure of Web 2.0 social media platforms (e.g., two-way, many-to-many, public /quasi-public) is criminologically important, as it is precisely these distinctive features that generate the conditions and opportunities for offending. For example, maintaining an extended online presence renders potential victims visible and accessible to prospective offenders; the public character of much social media use (the ability, on platforms such as Twitter, Pinterest, and YouTube, to "follow" and address any other users) amplifies the ability of offenders to both target others and share/disseminate prohibited content. It is therefore unsurprising that the emergence of social media has brought in its wake multiple vectors of criminal behavior. For analytical purposes it may be useful to adopt an initial two-fold classification of such offences, categorizations which can then be further sub-divided so as to isolate their distinctive features and characteristics. Initially, we can distinguish between person-centred and content-centred offences. The former comprise behaviors in which an offender targets particular individuals or groups who are subjected to a range of harms, such as those associated with bullying, harassment, hate speech, sexual predation, and fraud. The latter offences centre upon the online distribution of content that is legally prohibited (such as obscene words and images or incitement to commit violence) or content over which the sharer has no legal rights or ownership (such as copyrighted audio-visual materials). There are, in fact, a number of other typologies or classificatory schema that have been proposed as regarding illegal and harmful behavior on social media platforms. For example, Livingstone and Haddon (2009, p. 10) use a threefold classification that distinguishes between "content" (the individual is the recipient of a harmful communication), "contact"' (the individual participates in the offence as its target), and "conduct" (the individual particulates in the offence as the actor responsible for initiating the harmful behavior). While undoubtedly useful, this typology appears to conflate two different axes or criteria – the type of behavior (content vs. conduct) and the role of the individual in those behaviors (offender vs. victim). For present purposes, I deploy a simplified twofold distinction (content-oriented vs. person-oriented offences), while noting that for each of these offence types there will inevitably be both initiators (instigators, offenders) and targets (recipients, victims).

Furthermore, in respect to this broad typology, two specific issues are also worth noting. Firstly, there may occur offences that combine both inter-personal and content-centred elements, such as in cases where obscene (prohibited) speech is directed towards a specific individual who has been singled out for victimization (as in instances of stalking and harassment). Secondly, the kinds of behaviors noted here may be subject to differentiated treatment in law across different territories. Particularly notable in this regard is the divergence between many European jurisdictions and the United States. For example, European nations (such as the UK) have enacted legal prohibitions against speech that incites hatred against persons in relation to characteristics such as their race, ethnicity, religion, sexual orientation, and disability. In effect, in both offline and online settings, there are criminal sanctions in place intended to protect both groups and individuals from communication-based harms. By contrast, in the United States, First Amendment protections on free speech mean that the scope of laws against hateful content is far more narrowly circumscribed, being restricted to speech that amounts to "a serious and imminent threat of violence against identifiable persons, or directly incites others to commit specific criminal acts against those persons" (Yar, 2013, p. 103, emphasis in original). As one might imagine, this is a rather complex and detailed process, but one that makes sense and can be applied to the protection of any asset.

In light of both the range of different types of social media-based offences, and the cross-national variations in terms of legal prohibition, it is not possible in the present discussion to consider either the offences or their treatment in general terms. Rather, for the purposes of the present discussion, I will narrow the focus in terms of jurisdiction and offence type. I will largely confine my analysis to the

legal and criminal justice responses in the UK (or, to be more precise, England & Wales, as Scotland and Northern Ireland have their own distinctive legal systems). In terms of the offences considered here, the focus will fall upon (1) those offences that particularly target children and young persons, and (2) offences linked to hate, political extremism, radicalisation and the cultivation of support for terrorist causes and actions. This selection is based pragmatically upon the high profile that these problems have enjoyed of late in the UK, with a considerable degree of time and energy being directed to legislators, agents of law enforcement, charitable and other "third section" organisations, and mass media discussion. As such, these high-profile issues serve to exemplify the evolution of regulatory responses as well as the tensions and dilemmas that coalesce around these efforts. A second, subsidiary, criterion for focusing upon these offences is that it enables us to consider both person-centred and content-centred offence types. The nature and scope of these offences, as they relate to social media, will be outlined below, before discussion turns to regulatory, legal, and crime control responses.

Children and young persons (minors) have been amongst the most enthusiastic adopters of social media applications and the new technologies by which they can routinely be accessed. For example, a 2017 study by the UK media regulator Ofcom found that 74% of 12 to 15-year olds now have social media profiles, with 10% using such services to "live stream" themselves (i.e., sharing video of themselves and their activities online in real time) (Ofcom, 2017, p. 7). Despite most social media platforms setting a minimum user age requirement of 13 years, the study found that 23% of 8 to 11-year olds nevertheless have a social media account, with an uptake of 3% for those aged between 5 and 7 (Ofcom, 2017, p. 2). The prevalence, and hence potential vulnerability to victimization amongst minors, is further exacerbated by the means through which they access social media. With 83% of British 12 to 15-year olds now owning their own smartphones (and 55% having their own tablets), children's internet and social media usage has moved steadily toward more individualized modes, taking them further away from parental or other adult supervision while they are online (Ofcom, 2017). This ownership of mobile communication devices, alongside increased connectivity (e.g., using 4G data access and Wi-Fi hotspots), has also enabled children to access social media (and be reached in turn) across a variety of settings while away from home and family. Finally, the evidence indicates an increasing differentiation of young persons' choice of social media, with an uptake of newer platforms such as Snapchat, and a movement away from those platforms more likely to be used by their parents, such as Facebook (Ofcom, 2017, p. 7). Taken together, these developments create a context of heightened vulnerability as children increasingly access social media and do so in the absence of "capable guardians" such as parents who could supervise their communications and potentially intervene so as to prevent victimization (Cohen & Felson, 1979; Hollis, Felson, & Welsh, 2013). This dynamic also applies conversely to the possibilities that unsupervised internet and social media usage will result in minors engaging in harmful behavior towards other users, especially their peers.

The Ofcom report also offers some valuable insights into children's experiences of harmful and unwelcome content and behavior while online. For example, 45% of 12 to 15-year olds report having seen hateful content online ("something hateful ... directed at a particular group of people, based on, for instance, their gender, religion, disability, sexuality or gender identity"); 10% of the sample reported having seen "something of a sexual nature that made them feel uncomfortable;" and 25% say that they have been contacted online by someone they do not know (Ofcom, 2017, p. 14). We can also glean some insights into one of the most high-profile forms of online victimization involving minors, that of cyber-bullying. This term, which has rapidly gained currency in scholarly and public policy discussion, can be seen as a composite term that encompasses a range of behaviors in which individuals are targeted by malicious communications that denigrate, humiliate, threaten, and vilify the victims (Willard, 2007). The study found that 12% of 12 to 15-year olds reported having been bul-

lied in this way via social media; it is worth noting that this represents a significant increase on the findings from a previous iteration of the same survey in 2013, where the figure stood at 8% (Ofcom, 2013, p. 10). These findings align with previous studies of online bullying and harassment, including those from other countries, although the precise proportion of users experiencing victimization varies. For example, studies of cyber-bullying from the United States and Asia suggest that somewhere between 10% and 20% of children and young people experience such victimization in any given year (Holt, Bossler, &Seigfried-Spellar, 2015, pp. 214-216). While the extent and impact of such victimization will inevitably vary, there have been numerous cases in which concerted online bullying has been held responsible for serious mental health problems, self–harm, and even suicide (Hinduja & Patchin, 2010; Bauman, Toomey, & Walker, 2013). The other major area of concern regarding children's use of social media relates to child sexual abuse, including the targeting of minors with sexually explicit comments and suggestions, the solicitation of minors to perform sex acts, or to share explicit images online, as well as the "grooming" of minors as a precursor for offline contact abuse (Cano, Fernandez, & Alani, 2014; Wolak, Finkelhor, Mitchell, & Ybarra, 2008; Ybarra & Mitchell, 2008). One recent UK-based study of "technology assisted child sex abuse"(TA-CSA) found that electronic communication technologies "assisted the initiation, maintenance and escalation of abuse" in a number of ways: 1. by increasing the ease by which offenders can access victims; 2. by lowering inhibitions on behavior; 3. by creating feelings of powerlessness; 4. by facilitating emotional blackmail; 5. by facilitating image-related blackmail; and 6. diminishing victims' ability to recognise the abusive nature of the incidents (Hamilton-Giachritsis, Hanson, Whittle, & Beech, 2017, p. 6). Victims reported a range of negative impacts from the experience of online sexual abuse, including feelings of self-blame, depression, flashbacks, nightmares, panic attacks and anxiety, low self-esteem and feelings of worthlessness, self-harm, difficulties sleeping, eating disorders, and behavioral problems that affect ability to study (IHamilton-Giachritsis et al., 2017, p. 7). While studies suggest that the risk of sexual victimization is less commonplace than that associated with exposure to hateful and violent content or bullying, the seriousness of harms that ensue have ensured that the problem remains a high priority for those concerned with the online safeguarding of children (Livingstone & Haddon, 2013, p. 16).

Internet Governance and the Regulation and Control of Social Media

Having briefly outlined some of the key crime-related challenges arising from the growth of social media, we now turn to consider the kinds of responses these problems have engendered. However, these responses cannot be fully understood without appreciating the broader architecture of regulation and control that has developed around the internet, as the contours of social media regulation may be seen as a specific instance of these broader configurations. The internet's regulatory architecture has evolved in a manner that in many ways exemplifies the shift from "government" to "governance" in the execution of public policy. The concept of governance is tied to the rise of multi-party policy networks (spanning public, private, and voluntary sectors) which come together in order to effect functions of social-coordination or "steering" (Rhodes, 2007). Governance here refers to "changed conditions of ordered rule" (Rhodes, 1996, pp. 652-653) such that outcomes previously effected through centralized government actions are now achieved through different processes occurring outside the institutions of the nation state. This shift can be linked to a number of important drivers, including (1) the ideological and political reaction, since the 1970s, against top-down state interventionism, in favour of social functions and needs being met by "markets" and "civil society," and (2) a pragmatic need for new modes of coordination in conditions of ever-greater globalized complexity and interdependence (Crawford, 2006; Jessop, 2002). With regard to the internet, we can additionally note the influence of a "hacker ethic" amongst the engineers, scientists, and coders who pioneered and popularized computer-

mediated communication, a value orientation that emphasized individual freedom and suspicion of excessive state control (Steinmetz, 2016). As such, the development of internet regulation has featured an "arms-length" relationship between state agencies and other actors that are involved in managing and "policing" online activity on a daily basis.

We see how the ordering of the internet (especially in relation to crime control and prevention) is affected through a variety of instruments, wielded by a range of non-state as well as state-based actors. These include laws (criminal and civil), the setting of regulatory standards and protocols, codes of practice, as well as a range of technical measures or "fixes," such as digital rights management (DRM) tools to prevent unauthorized copying of protected media content such as movies and software, or user authentication tools to control and limit access to secure systems like those for online banking. We can suggest two main aspects of these efforts to institute internet governance, namely those associated with the creation of laws, regulations, and standards, and those associated with their implementation or enforcement. Each are considered below, with particular attention to those measures most relevant to controlling crime on social media.

When it comes to "rule setting" (what is permissible to do or not do while online), falls in the first instance to states and inter-state actors (the latter includes transnational political unions such as the EU, intergovernmental organizations such as the UN, WTO, Council of Europe, International Communications Union (ICU)) and so on). States either individually or collectively through treaties and the formulation of international public law set in place legally-sanctioned prohibitions around online conduct. Thus, in respect to the kinds of online crime issues considered here, we have seen sequential legislative innovation over the past few decades in the UK, as elsewhere. This has involved creation of new laws that cover, in whole or part, internet-related offences. These include the Computer Misuse Act (1990), which criminalizes (a) acts such as unauthorized access to, and interference with, computer systems, and (b) the engagement in such acts that facilitate further offences (Macewan, 2008). Other laws address specific kinds of undesirable online behavior. With respect to the aforementioned problem of communication containing sexualized and pornographic representations of minors, the reach of existing obscenity and child protection laws (such as the Obscene Publications Act (1959) and the Protection of Children Act (1978)) has been incrementally extended and enhanced through further legislation. For example, concern about animated (cartoon) and digitally-generated sexual representations of minors on the internet resulted in legislation that prohibits the production, possession, and transmission of such material under provisions in the Criminal Justice and Public Order Act (1984), the Criminal Justice and Immigration Act (2008) and the Coroners and Justice Act (2009) (Akdeniz, 2008; Gillespie, 2011, 2018).

With regard to online child sex abuse, provisions include those of creating an offence of "sexual grooming" under the Sexual Offences Act (2003). Most recently, S.67 of the Serious Crimes Act (2015) creates an offence of "sexual communication with a child", for which conviction carries a custodial sentence of up to two years (Ministry of Justice, 2017). Non-sexually oriented forms of online abuse, such as that associated with cyber-bullying, have been covered by provisions such as those in the Malicious Communications Act (1988), the Protection from Harassment Act (1997), the Communications Act (2003), and the Defamation Act (2013) (El Asam & Samara, 2016). The other area of intense concern already identified, namely that of extremist and terrorist-related content and interaction on social media has likewise been subjected to statutory measures and criminal law sanctions. For example, the Terrorism Act (2000) includes measures to criminalize communications that contain terrorist threats, express or invite support for proscribed groups, solicitation of money or other resources for the purposes of terrorism, collecting or making "a record of information of a kind likely to be useful to a person

committing or preparing an act of terrorism," and inciting others to commit acts of terrorism outside the UK. The 2006 Terrorism Act goes further and prohibits the "glorification" and "encouragement of terrorism,"' which includes communications shared via the internet and social media (Edwards et al., 2010, pp. 421-423). In the case of R v Gul, the accused was found guilty and sentenced to five years imprisonment for "encouragement of terrorism" after posting on YouTube and other online platforms videos showing attacks by the Taliban, al-Qaeda, and other proscribed groups (Coco, 2013). Most recently, the UK government has propozed further tightening of the law through measures that could lead those who "repeatedly view terrorist content online" to 15 years imprisonment, in an attempt to counter online "radicalisation" of those who go on to commit attacks (Travis, 2017).

Beyond the prohibitions set out in law, the regulation of online behavior has been affected by rule setting emanating from those who provide online communication platforms and applications. For example, the social networking platform, Facebook, makes as a precondition of use the adherence to predefined "community standards" covering issues such as graphic violence, nudity and sexualized imagery, threats and bullying, and the procurement and sale of prohibited goods and sexual services (Facebook, 2017a). Similarly, Twitter's "user agreement"'prohibits behavior including IP violations, posting "abusive and hateful content," unauthorised sharing of someone else's private information or "intimate media" (e.g., nude images, "sex tapes" etc.), and distribution of spam and malware (Twitter, 2017a). Other popular social media platforms, such as Instagram and Snapchat, also institute comparable constraints upon users. The parameters of such rules, which act as boundary-markers for delineating the acceptable from the harmful, dangerous or illicit – typically enact formal legal sanctions and prohibitions but can also exceed them by drawing upon broader social and cultural norms about acceptable and unacceptable content and behavior. For example, with Facebook's rules against graphic nudity and sexualized content, which are oriented not by laws around obscenity but by user and community concerns that minors and other vulnerable persons not be exposed to content deemed "inappropriate" or unwelcome. Other non-state actors (such as NGOs, charities, pressure groups, and campaigning organisations) may also play a decisive role in negotiating and shaping such rules and sanctions. A notable example is provided by the case of child protection charities that seek to advise and influence service- and content-providers so as to prevent sharing of sexualized images of minors and inappropriate contact between adult users and children. In sum, what we see here is a dynamic system of rule-making that involves numerous state and non-state actors in setting the parameters for the social control of online behavior. We shall return shortly to consider the effectiveness and limitations of these rules at the level of their implementation.

The enforcement of social control measures related to crime on social media can likewise be identified through a range of multi-party or "mixed" (state, quasi-state, NGO, public and private) governance configurations. Part of the day-to-day business of "policing the Internet" remains with traditional law-enforcement actors located at both state (national) and sub-state (local) levels. Taking again the example of England and Wales, all 43 of its territorially-based police forces have some kind of specialist provision for tackling computer-related and internet-based offences, albeit organized in varying configurations. At the national level, the National Crime Agency now incorporates what was previously the Police Central e-crime Unit (PCeU), and takes a lead role in tackling cybercrime alongside related problems such as organised crime and economic crime. Operationally, it is clear that the police (and in relation to terrorist offences, the security services) take an active role in monitoring social media traffic. Such surveillance can be used for both intelligence gathering and monitoring in relation to suspect groups and individuals, and retroactively for investigative evidence gathering in relation to reported crimes (Schneider, 2016; Trottier, 2012, 2015). The number of such offences being brought to the attention of police appears to be rising. Reports of online child sex abuse referred to London's

Metropolitan Police increased 700% between 2004 and 2017 (Bowcott, 2018). Across England & Wales more broadly, in 2016-17 there were 5,653 police recorded incidents of sexual offences against children containing an "online element," an increase of 1,744 on the figure for 2015-16 (Bowcott, 2018). Such figures are indicative of increasing levels of policing activity directed at victimisation via communications media. However, the policing of social media by state agencies sits alongside the involvement of private (non-state) actors in social control efforts. It is frequently noted by cybercrime researchers that the capacities of public law enforcement are outstripped by the scale and complexity of the online environment and there is consequently a significant deficit when it comes to the ability of criminal justice systems to respond to cyber-criminal conduct (Wall, 2007). Consequently, tied to the behavioral strictures set-out in social media platforms' rules and "community standards" we find ongoing enforcement efforts. User content and behavior on such platforms are routinely assessed by teams of paid moderators. Facebook currently employs some 7,500 full-time staff in this capacity and breach of the rules will result in content removal and even expulsion of offenders from the platform (Kuchler, 2017).

The efforts of human moderators are increasingly supplemented by automated (machine learning) systems that use algorithms to identify and remove suspect content on the platforms (CDT, 2017). However, figures about the precise number of such "takedowns'" of violating content are limited and incomplete. For example, while Facebook declares in its bi-annual Transparency Report the number of content removals on grounds of intellectual property (copyright and trademark) violations, it does not provide comparable figures for takedowns of other kinds of material, such as pornographic and obscene content, hate speech, incitement to violence, and the like (Facebook, 2017b). Similarly, while Twitter produces figures every six months relating to removal requests received and acted upon (14,120 and 3,032 respectively for the first half of 2016), and also enumerate such requests by country of origin, it does not offer any insights about the grounds for complaint or removal, as in which laws or Terms of Service the offending content is held to have violated (Twitter, 2017b). None of the major social media platforms appear to publish data on the numbers of accounts that are suspended or disabled for violating rules addressing offensive content and behavior. Consequently, despite the platforms' very public assurances about their commitment to regulating and removing illegal and otherwise problematic content and the individuals or groups responsible for it, the extent of such control remains unknown. However, as we shall consider in the next section, social media providers have recently come under concerted criticism for their alleged failure to effectively "police" their own domains, leading to calls for more coercive measures to compel action.

Failure to Regulate and the Shifting Public Policy Response to Crime on Social Media

In 2007, the UK government commissioned the so-called "Byron Review," led by clinical psychologist Tanya Byron, to assess the development of young people's internet use (especially social media), to identify the risks and harms to which they are exposed online, and to make recommendations about how internet use could best be regulated and policed (in the broad sense of the term) so as to ameliorate those harms. Reporting a year later (under the title "Safer Children in a Digital World"), the review highlighted the need for collaborative solutions involving "families, industry, government and others in the public and third sectors'" and for "better self-regulation" on the part of online platforms (Byron, 2008, p. 2). Specifically, it recommended that "the relevant industries should develop an independently monitored voluntary code of practice on the moderation of user generated content, including making specific commitments on take-down times" (Byron, 2008, p. 6). In response, the government established the UK Council for Internet Child Safety, that brought together more than 200 organisations drawn

from government, law enforcement, academia, the internet industry, media, and charities, with a view to delivering on the Byron Review's recommendations about enhancing online child safety.

Moving forward almost a decade, we can assess the perceived failures in effectively addressing the challenges identified by Byron or implementing properly the report's recommendations, alongside progress (or lack thereof) in tackling other forms of prohibited content and behavior on social media. In 2017, the MP Alex Chalk led an inquiry into cyber-bullying, conducted by the children's charities, The Children's Society and YoungMinds. Based upon a combination of survey data about children's experiences of internet and social media use, expert evidence, testimony from new media companies, and written submissions from third sector organisations, Chalk's report found that:

(1) Despite social media platforms' stated age restrictions on users, some 61% of young people created their first accounts at the age of 12 or younger.

(2) Cyber-bullying is a growing problem which results in tangible and serious mental health problems.

(3) Social media platforms' responses to the problem continue to be "inconsistent and inadequate."

Consequently, the inquiry recommended that social media companies be required to more effectively monitor and control access according to age, that there be industry-wide standards requiring companies to respond to and resolve reports of abuse within 24 hours, that they effectively train their content moderators in relation to child-safety and mental health, and that companies use algorithmic monitoring tools to better identify incidents and patterns of offending behavior (The Children's Society & YoungMinds, 2017, p. 62). In other words, it identified that the kinds of recommendations regarding social media self-regulation proposed by the Byron Report had tangibly failed to be implemented, leading to an exacerbation rather than amelioration of the problem.

Correspondingly, it argued that government needs to take a more directive role in ensuring accountability for child safety from social media companies through requiring, recording and reporting data on the numbers of breaches of their "community standards," recording the nature of the incidents reported, recording the speed of response or resolution (e.g., take-downs of offending content or suspending offenders' accounts), instituting mandatory annual transparency reporting, and the creation of an independent arbitration service for dealing with user complaints about unsatisfactory responses to incidents on the part of the companies (The Children's Society & YoungMinds, 2017). What is significant here is a discursive shift in the language of social control from industry self-regulation and user "responsibilization" to a more active and potentially coercive role for the state regarding the content and interactions taking place online. The growing criticisms of social media platforms and calls for more robust and statutory regulation were further amplified by sometimes lurid press reports about the unchecked levels of child exploitation online. Recent headlines give a flavour of the growing clamour and recrimination aimed at social media. Some examples include "Social media firms 'failing' to tackle cyber-bullying" (BBC News 2018a), "Facebook rife with grooming" (Alton Herald, 2018), and "Children as young as 7 are being groomed online" (McDonald, 2018).

In 2017, the UK Government published provisional proposals (a so-called Green Paper) as part of its latest Internet Safety Strategy, but this largely reiterated its previous calls for a self-regulatory approach through "codes of practice," in combination with user education and "responsibilization" in terms of parents deploying technological solutions such as filters and content locks. Tanya Byron weighed in by arguing that "social media companies have let children down because we have let them play by their own rules," and that the government response in the Green Paper has been "half-baked and insulting" (quoted in Woolcock, 2017). Alongside other children's charities and child safety campaigners,

she called instead for an independent regulator of social media with the power to compel companies to institute protective measures, and the ability to punish failure to perform adequately through the use of fines (NSPCC, 2017).

In tandem with the above-discussed shortcomings in relation to child-oriented abuses, a parallel line of criticism has emerged around the proliferation of hate speech and extremist political communication online. Such evaluations about the shortcomings of social media have emerged not just from NGOs, academic observers, and the press, but also from within the legislature. Most notably, the UK Parliament's Home Affair Select Committee published, in 2017, its report "Hate Crime: Abuse, Hate and Extremism Online." Having considered evidence about the circulation of Islamophobic, misogynistic, far-right extremist and terrorism-related content on social media, the Committee concluded that:

> . . . there is a great deal of evidence that these platforms are being used to spread hate, abuse and extremism. That trend continues to grow at an alarming rate but it remains unchecked and, even where it is illegal, largely unpoliced (Home Affairs Select Committee, 2017, p. 4).

They went on to condemn "the laissez-faire approach that many social media companies have taken to moderating extremist content on their platforms," and in the case of Google even suggested that it has "profited from hatred and has allowed itself to be a platform from which extremists have generated revenue" (Home Affairs Select Committee, 2017, p. 10). They point to what is seen as a consistent pattern of failure to remove such illegal content, even when alerted to its presence by users and others, and note the far more prompt and agile take-down responses when protecting commercial interests in relation to copyright violations (Home Affairs Select Committee, 2017, p. 4). In a significant reorientation, the Committee has urged that government "should now assess whether the continued publication of illegal material and the failure to take reasonable steps to identify or remove it is in breach of the law" (Home Affairs Select Committee, 2017, p. 12). In other words, legal liability for the breach of laws prohibiting hate speech and support for terrorism would under such proposal lie not just with the users who post and share the offending content, but also by extension with the content's "publishers," should they fail to remove it in timely fashion as dictated by law. Finally, such liability should further be underpinned by "a system of escalating sanctions to include meaningful fines for social media companies which fail to remove illegal content within a strict timeframe" (Home Affairs Select Committee, 2017, p. 22).

Alongside the calls for statutory regulation of social media in relation to child-centred offences considered earlier, the stance of the Select Committee is testimony to what could be a landmark moment in relation to crime control in social media - the shift from a self-regulatory model of governance in which the state acts as coordinator, facilitator and guide, to one in which the force and sanctions available in law are brought to bear directly and concertedly upon social media platforms so as to compel more effective content control. It is also important to note, as related to the extension of liability to social media companies, the potential impact of the recently-adopted General Data Protection Directive (GDPR) of the EU. The GDPR as a whole, places additional responsibilities upon those who "control" and "process" citizens' personal data, and places particular emphasis upon issues of consent and care when handling data related to minors (those under the age of 16). Insofar as the kinds of offences considered here may entail data breaches (i.e.,"a breach of security leading to the destruction, loss, alteration, unauthorised disclosure of, or access to, personal data") (ICO, 2017, p. 88) the social media platform will be liable and the user will have the right to compensation. Moreover, the GDPR makes provision of administrative fines of up to £20 million for non-compliance (Burgess, 2018), thereby placing considerable pressure upon the platforms to protect users' data from misappropriation by

malicious actors. These changes, in tandem with the kinds of proposed measures already considered above, may add further pressure for a more directive and state-centric model of control in the UK.

At the time of writing this, the UK government also appears to be shifting its stance away from the self-regulatory approach favoured in its own Internet Safety Strategy of 2017, and toward a more legally-driven and law-enforcement based approach to social media. In April 2018, the Secretary of State for Health, Jeremy Hunt, decried social media companies' responses to previous calls to action as "extremely limited" and, writing directly to those companies, accused them of "collectively turning a blind eye to a whole generation of children being exposed to the harmful emotional side effects of social media" (quoted in BBC News, 2018). This stance appears to pave the way for potential new legislation that would shift responsibility for illegal content to the platforms that have heretofore enjoyed only limited liability as "intermediaries." Such a change would only become possible upon the "Brexit," when EU's E-Commerce Directive, which exempts online service providers from a general duty to monitor for and remove illegal content, could be set aside. Indeed, another Parliamentary Committee (The Committee on Standards in Public Life) has already recommended such a shift in its July 2017 report (Committee on Standards in Public Life, 2017, pp. 13-14; Out-Law.com, 2017). Should such a shift indeed occur in the coming years, it would no doubt prove controversial, as it has both commercial implications for the companies themselves (in terms of accruing potentially massive financial liabilities for failures to remove offending content), and in terms of the potential "chilling effects" on online free speech, as such laws would incentivize companies to strictly limit the kinds of content permitted of users.

Finally, we need to consider the challenges and dilemmas faced by social media companies themselves. The criticisms of them certainly have merit, and they can justifiably be accused of a failure to properly discharge their social responsibilities to mitigate the ill-effects of the communications platforms which have made them extremely successful; Facebook alone generated $40.5 billion in revenue in 2017, while Twitter generated $732 million in revenue and $91 million in profit in the last quarter of the year (Roettgers 2018, Wagner 2018). However, in their own defence, the companies point out that the sheer scale of the challenge makes comprehensive and rapid take-down of offending content an uphill struggle, even when employing a battalion of moderators and machine-learning algorithms to facilitate automatic identification of such content. For example, in 2013 Facebook alone saw 350 million photos being uploaded by users every day (Smith 2013), any of which may potentially be in breach of legal or other prohibitions. Moreover, there is ample evidence that automatic content identification and blocking software generates a significant number of "false positives," thereby unintentionally blocking or removing content that is entirely legal and/or in line with the platform's "community standards" (Duarte et al. 2018).

This problem draws our attention to a broader tension around such companies' content control practices – while on the one hand they are criticized for a lack or vigour in monitoring and removing problematic content, they are castigated from other quarters for being overly-censorious and thereby stifling discussion of controversial social and political issues (see Wong et al. 2017). This ultimately brings us to the conflict between crime control and free speech, with the advocates of (digital) liberties urging social media to resist government-mandated intrusion and control over users' communication practices (Levinson, 2017). An alternative to legislatively-mandated liabilities and punitive sanctions may be to further emphasize the role of user-centred protections including greater education and awareness, so as to prevent victimization rather than reacting to it via the criminal justice system after the fact (LaRose et al. 2008; Palfrey et al. 2010; Shillair et al. 2015). However, advocates of a more directive role for the state might point out the relative failure of past education- and prevention-oriented initi-

atives to significantly dent levels of online victimization. In this dynamic, volatile and rapidly shifting arena, the shape of a new architecture of crime control on social media remains, as yet, to be seen.

Conclusion

The rapid development and popular adoption of social networking platforms has, as with other internet-related practices, brought to the foreground a range of crime problems that have an increasingly negative impact. These impact individual application users (both children and adults), as well as broader consequences for social groups, minorities, national security and political stability. The confluence of a series of high-profile cases has placed a critical spotlight upon the mechanisms through which social media are regulated, and their content and users' behavior are policed. The founding model focused upon self-regulation and voluntary content control by media companies (assisted by the state and third-party actors who offer additional support through policy framing, voluntary initiatives, and educational outreach). However, the perception of sequential failures to effectively address online crime and harassment, alongside further scandals (such as those related to "fake news," electoral manipulation by foreign agents, and misuse of users' personal data) has created in the UK, and elsewhere, a discursive shift. We see now the first steps in what may be a move to install more rigorous and mandatory state-sanctioned measures that make use of direct institutional oversight, "responsibilization" of media platforms, and the potential use of punitive and criminal law sanctions against such companies so as to compel action. Such steps, justified by the necessity to tackle increasingly troubling levels of victimisation, place social media at the centre of a conflict between users' rights to free speech and the demands of crime control, as well as presenting daunting practical challenges around how unprecedented flows of mediated communication and interaction can be effectively monitored and managed. The regulatory architecture that will emerge in the UK in response to these tensions and dilemmas may well have far-reaching consequences for cybercrime control over many years to come.

References

Akdeniz, Y. (2008). *Internet child pornography and the Law: National and International Responses*. London: Routledge.

Alton Herald (2018, April 19). *Facebook rife with grooming*. Retrieved from http://www.altonherald.com/article.cfm?id=126531&headline=Facebook%20rife%20with%20grooming.

Bauman, S., Toomey, R. B., & Walker, J. L. (2013). Associations among bullying, cyberbullying, and suicide in high school students. *Journal of Adolescence, 36*(2), 341-350.

BBC News (2018a, February 26). *Social media firms 'failing' to tackle cyber-bullying*. Retrieved from http://www.bbc.co.uk/news/technology-43197937.

BBC News (2018b, April 22). *Jeremy Hunt threatens social media with new child-protection laws*. Retrieved from http://www.bbc.co.uk/news/uk-43853678.

Bird, S. E. (2011). Are we all produsers now? Convergence and media audience practices. *Cultural Studies, 25*(4-5), 502-516.

Bowcott, O. (2018, Jan 22). Online child sex abuse referred to met increased by 700% since 2014. *The Guardian*. Retrieved from urlhttps://www.theguardian.com/uk-news/2018/jan/22/online-child-sex-abuse-referred-to-met-increased-by-700-since-2014.

Boyd, D. & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230.

Byron, T. (2008). textitSafer children in a digital world: The report of the byron review. Nottingham: Department for Children, Schools and Families and the Department for Culture, Media and Sport.

Burgess, M. (2018, June 4). What is GDPR? The summary guide to GDPR compliance in the UK. *Wired.* Retrieved from http://www.wired.co.uk/article/what-is-gdpr-uk-eu-legislation-compliance-summary-fines-2018.

Cano, A. E., Fernandez, M., & Alani, H. (2014, November 11). *Detecting child grooming behaviour patterns on social media.* In L. M. Aiello & D. McFarland (Eds.), *International conference on social informatics* (pp. 412-427).

Coco, A. (2013). The Mark of cain: The crime of terrorism in times of armed conflict as interpreted by the court of appeal of England and Wales in R v. Mohammed Gul. *Journal of International Criminal Justice, 11*(2), 425-440.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588-608.

Committee on Standards in Public Life (2017). *Intimidation in public life: A review by the committee on standards in public life*. London: HMSO.

Cormode, G. & Krishnamurthy, B. (2008). Key differences between Web 1.0 and Web 2.0. *First Monday, 13*(6). http://firstmonday.org/ojs/index.php/fm/article/view/2125/1972.

Crawford, A. (2006). Networked governance and the post-regulatory state? Steering, rowing and anchoring the provision of policing and security. *Theoretical criminology, 10*(4), 449-479.

Duarte, N., Llansó, E. & Loup, A. (2018). Mixed Messages? The limits of automated social media content analysis. Paper presented at the 2018 Conference on Fairness, Accountability and Transparency.

Edwards, L., Rauhofer, J. & Yar, M. (2010). Recent developments in UK cybercrime law. In Y. Jewkes & M. Yar (Eds.), *Handbook of Internet Crime* (pp. 413-436). Cullompton: Willan.

El Asam, A., & Samara, M. (2016). Cyberbullying and the law: A review of psychological and legal challenges. *Computers in Human Behavior, 65*, 127-141.

Facebook. (2017a). Community Standards. Retrieved from https://en-gb.facebook.com/communitystandards.

Facebook. (2017b). Transparency report: Intellectual property. Retrieved from https://transparency.facebook.com/intellectual_property/.

Gillespie, A. A. (2011). *Child pornography: Law and policy*. Abingdon & New York: Routledge.

Gillespie, A. A. (2018). Child pornography. *Information & Communications Technology Law, 27*(1), 30-54.

Hall, K. (2017, April 17). More than 87m Facebook profiles farmed, says second ex-Cambridge. *The Register*. Retrieved from https://www.theregister.co.uk/2018/04/17/former_cambridge_analytica_staffer_brittany_kaiser_dcms_committee_evidence/.

Hamilton-Giachritsis, C., Hanson, E., Whittle, H. C., & Beech, A. R. (2017). *Everyone deserves to be happy and safe. A mixed methods study exploring how online and offline child sexual abuse impact young people and how professionals respond to it*. London: NSPCC.

Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research, 14*(3), 206-221.

Hollis, M. E., Felson, M., & Welsh, B. C. (2013). The capable guardian in routine activities theory: A theoretical and conceptual reappraisal. *Crime Prevention and Community Safety, 15*(1), 65-79.

Holt, T. J., Bossler, A. M., & Seigfried-Spellar, K. C. (2015) *Cybercrime and digital forensics: An introduction*. New York, NY: Routledge.

Jessop, B. (2002). Liberalism, neoliberalism, and urban governance: A state–theoretical perspective. *Antipode, 34*(3), 452-472.

Kuchler, H. (2017, May 3). Facebook to hire 3,000 more moderators to check content. *Financial Times*.

ICO (Information Commissioner's Office). (2017). Overview of the General Data Protection Regulation (GDPR). Retrieved from https://ico.org.uk/media/for-organisations/data-protection-reform/overview-of-the-gdpr-1-13.pdf.

LaRose, R., Rifon, N. J., & Enbody, R. (2008). Promoting personal responsibility for internet safety. *Communications of the ACM, 51*(3), 71-76.

Levinson, P. (2017, November 28). Government regulation of social media would be a 'cure' far worse than the disease. *The Conversation*. Retrieved from ttp://theconversation.com/government-regulation-of-social-media-would-be-a-cure-far-worse-than-the-disease-86911.

Livingstone, S., & Haddon, L. (2009). *EU kids online: Final report 2009*. London: EU Kids Online Network.

Macewan, N. F. (2008). The Computer Misuse Act 1990: lessons from its past and predictions for its future. *Criminal Law Review, 12*, 955-967.

McDonald, G. (2018, April 18). Children as young as 7 are being groomed online. *The Plymouth Herald*. Retrieved from https://www.plymouthherald.co.uk/news/plymouth-news/children-young-7-being-groomed-1464465.

Ministry of Justice (2017). *Sexual communication with a child - implementation of s.67 of the Serious Crime Act 2015*. Circular No. 2017/01. London: HMSO.

NSPCC (2017, April 27). Social media sites failing to protect children. *NSPCC*. Retrieved from https://www.nspcc.org.uk/what-we-do/news-opinion/social-media-sites-failing-protect-children/.

Ofcom (2013). *Children and parents: Media use and attitudes report*. London: Ofcom.

Ofcom (2017). *Children and parents: Media use and attitudes report*. London: Ofcom.

O'Reilly, T. (2010). What is Web 2.0? Design patterns and business models for the next generation of software. In H. M., Donelan, K. Kear, & M. Ramage (Eds.), *Online communication and collaboration: A reader* (pp. 225-235). Abingdon: Routledge.

Outlaw.com (2018, December 13). UK government called on to shift liability for illegal online content to social media companies. *Outlaw.com*. Retrieved from https://www.out-law.com/en/articles/2017/dec

ember/uk-government-called-on-to-shift-liability-for-illegal-online-content-online-to-social-media-companies1/.

Palfrey, J., Boyd, D., & Sacco, D. (2010). *Enhancing child safety and online technologies: Final report of the Internet safety technical task force*. Durham, NC: Carolina Academic Press.

Rhodes, R.A.W. (1996). The new governance: Governing without government, *Political Studies, 44*, 652-657.

Rhodes, R.A.W. (2007). Understanding governance: Ten years on, *Organization Studies, 28*(8), 1243-1264.

Ritzer, G., Dean, P., & Jurgenson, N. (2012). The coming of age of the prosumer. *American Behavioral Scientist, 56*(4), 379-398.

Roettgers, J. (2018, January 31). Facebook says it's cutting down on viral videos as 2017 revenue Tops $40 billion. *Variety*. Retrieved from http://variety.com/2018/digital/news/facebook-q4-2017-earnings-1202683184.

Shillair, R., Cotten, S. R., Tsai, H. Y. S., Alhabash, S., LaRose, R., & Rifon, N. J. (2015). Online safety begins with you and me: Convincing Internet users to protect themselves. *Computers in Human Behavior, 48*, 199-207.

Schneider, C. J. (2016). *Policing and social media: Social control in an era of new media*. Lanham, MA: Lexington Books.

Smith, C. (2013, May 18). Facebook users are uploading 350 million new photos each day. *Business Insider.* Retrieved from http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9?IR=T.

Steinmetz, K. F. (2016). *Hacked: A radical approach to hacker culture and crime*. New York, NY: NYU Press.

Swaine, J. & Bennetts, M. (2018, February 17). Mueller charges 13 Russians with interfering in US election to help Trump. *The Guardian*. Retrieved from https://www.theguardian.com/us-news/2018/feb/16/robert-mueller-russians-charged-election.

The Children's Society & YoungMinds (2017). *Safety net: Cyberbullying's impact on young people's mental health: Inquiry report*. London: The Children's Society.

Travis, A. (2017, October 3). Amber Rudd: Viewers of online terrorist material face 15 years in jail. *The Guardian*. Retrieved from https://www.theguardian.com/uk-news/2017/oct/03/amber-rudd-viewers-of-online-terrorist-material-face-15-years-in-jail.

Trottier, D. (2012). Policing social media. *Canadian Review of Sociology / Revue canadienne de sociologie, 49*(4), 411-425.

Trottier, D. (2015). Open source intelligence, social media and law enforcement: Visions, constraints and critiques. *European Journal of Cultural Studies, 18*(4-5), 530-547.

Twitter. (2017a). The Twitter rules. *Twitter*. Retrieved from https://support.twitter.com/articles/18311.

Twitter. (2017b). Removal Requests. *Twitter*. Retrieved from https://transparency.twitter.com/en/removal-requests.html.

Ybarra, M. L., & Mitchell, K. J. (2008). How risky are social networking sites? A comparison of places online where youth sexual solicitation and harassment occurs. *Pediatrics, 121*(2), 350-357.

Wagner, K. (2018, February 8). Twitter just reported its first profitable quarter ever, but didn't add any new users in Q4. *Recode*. Retrieved from https://www.recode.net/2018/2/8/16989834/twitter-q4-2018-earnings-revenue-jack-dorsey.

Willard, N. E. (2007). *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats and distress*. Champaign, IL: Research Press.

Wolak, J., Finkelhor, D., Mitchell, K. J., & Ybarra, M. L. (2008). Online "predators" and their victims: Myths, realities, and implications for prevention and treatment. *American Psychologist, 63*(2), 111.

Wong, J. C., Safi, M., Rahman, S. A. (2017, September 20). Facebook bans Rohingya group's posts as minority faces 'ethnic cleansing', *The Guardian*. Retrieved from https://www.theguardian.com/technology/2017/sep/20/facebook-rohingya-muslims-myanmar.

Woolcock, N. (2017, November 20). Social media companies have let children down. *The Sunday Times*. Retrieved from https://www.thetimes.co.uk/article/social-media-companies-have-let-children-down-22wk6hw82#.