Social media corporations as actors of counter-terrorism

Marguerite Borelli

***

This is the accepted version of a manuscript published as Borelli, M. 2021. Social media corporations as actors of counter-terrorism. *New Media & Society*. doi: 10.1177/14614448211035121

***

Social media corporations as actors of counter-terrorism

Marguerite Borelli (CARISM, Université Paris II Panthéon-Assas)

Abstract

This article discusses the role of giant social media corporations Facebook, Google (YouTube), and Twitter in counter-terrorism and countering violent extremisms (CT/CVE). Based on a qualitative investigation mobilizing corporate communications as well as a collection of interviews with European stakeholders, it argues that these firms have become actors in this policy area of what is traditionally considered high politics, through their fundamental role in establishing and enforcing the nascent global governance regime on terrorist communications. Since the emergence of ISIS, the studied firms have displayed agency and creativity in their appropriation of this new responsibility, effectively going beyond what was legally required of them. After contextualizing and questioning their involvement, motivated by terrorist exploitation of their services, reputational pressure and the threat of legislation; the article provides an analysis of the firms' self-regulated commitment to CT/CVE through policymaking, content moderation, human resources and private multilateralism.

Introduction

As social media corporations grow more confident in appropriating their global "court-of-law-like powers" on freedom of expression (Musiani et al, 2016: 17; Badouard, 2020; Klonick, 2018), scholarly interest in their role in global politics is fast growing. Indeed, their reach expands to even high politics as some platforms have become global battlefields, bestowing "nation-sized political responsibilities" upon "carefree engineers uninterested in politics" (Singer and Brooking, 2018: 223; Arpagian in Musiani et al., 2016; de Goede, 2018). The involvement of private actors in security is not a new phenomenon, however neoliberal reconfigurations of state power have accelerated the tendency towards a "privatization of security" (Abrahamsen and Leander, 2016). In fact, a characteristic trend of the post-9/11 world is the emergence of "security assemblages" i.e. "transnational structures and networks in which a range of different actors and normativities interact, cooperate and compete to produce new institutions, practices and forms of deterritorialized security governance" (Abrahamsen and Williams, 2010: 90).

In the face of transnational and homegrown terrorism, governments are increasingly supplementing their hard counterterrorism (CT) strategies with softer countering violent extremism (CVE) measures[1], which involve a large range of actors from civil society alongside law enforcement and the military (Beutel and Weinberger, 2016; Ganesh and Bright, 2020; de Goede, 2018). Terrorism and violent extremisms are thus increasingly governed through security assemblages, or "fluctuating arrangements of networks of state, corporate and other

voluntary actors" (Tréguer, 2019: 148). Within this move toward whole-of-society approaches, the largest and most established social media corporations Google (Youtube), Facebook (Facebook, Instagram) and Twitter have been designated as key partners[2], often with the underlying assumption that their services contribute to (violent) radicalization (Hoskins and O'Loughlin, 2010; Ritzmann, 2017). Terrorist contents online are hence considered a priority by public and private stakeholders alike, because of their suspected real-world implications in terms of physical safety and security. Indeed, according to a Public Policy representative from Google : "in terrorism you have an extra sense of responsibility to get it right, because the real-world consequences of wrong choices are profound". Because terrorist content generates a consensus amongst stakeholders, its governance is more advanced, and studying it may provide an indirect way of obtaining information on the general direction of content regulation. Initiatives on hate speech and disinformation, for instance, often take past CT/CVE efforts as a point of reference and emerge through the same channels that were established to tackle terrorism, such as the EU Internet Forum (Coche, 2018)[3]. That is why, amidst the variety of content issues social media firms are faced with today, terrorism is a good place to start investigating their nascent political power in the security field.

This article argues that Google, Facebook and Twitter have become actors in the policy area of CT/CVE, through their fundamental role in establishing and enforcing the global governance regime on terrorist communications, i.e. the set of "implicit or explicit principles, norms, rules, and decision-making procedures around which actors' expectations converge" in this area of global politics (Krasner 1982:186). It provides an empirically-informed overview of how these large firms "reluctantly learn to see the world through a security lens" (de Goede 2018: 26), appropriating this new role, all the while balancing security concerns with economic imperatives and reputational incentives. After a brief presentation of the study's research design, the first part contextualizes and questions the involvement of these firms in CT/CVE, looking at how terrorist exploitation of their services, public pressure and the threat of legislation forced them into action. The second part then provides a review of the four main areas of their self-regulated CT/CVE activities, namely private policymaking, moderation, human resources and private multilateralism. Specific emphasis is placed on the agency and creativity displayed by the platforms in their self-commitment, as creativity in the form of novel, experimental practices can be understood as a means for firms to exercise corporate power (Hoffman, 2013; Malhotra et al., 2019).

Research design

The findings presented in this article are drawn from a qualitative investigation conducted between May 2018 and March 2019. This period saw multiple relevant developments, including EU negotiations around the Regulation on preventing the dissemination of terrorist contents online (TERREG), and the Christchurch attack.

Open sources used include a) an archive of Google, Facebook and Twitter corporate blog posts[4] up to February 2019 discussing CT/CVE at various lengths (68 posts, 66 000+ words); b) parts of their Terms of Service (specifically the Twitter Rules, Community Standards for Facebook, and Community Guidelines for YouTube); and c) public video archives from US congressional hearings of high-ranking company officials on content issues. To complement this publicly available information, 14 semi-structured interviews were conducted with relevant actors from France, Germany and the UK[5]. Lastly, additional data was collected at the 2018 Paris Peace

Forum, where Google and Facebook presented their CT/CVE efforts as participants and co-sponsors of the event.

1. Caught in the crossfire between terrorists and public authorities

Google, Facebook and Twitter could well have continued to frame their minimal involvement in CT/CVE through their "mere conduits" legal status, a wide interpretation of freedom of speech, and the defense of their flagging system. YouTube did this for instance in 2008 when it was accused of aiding and abetting al Qaeda by US Senator Joe Lieberman (Gillespie, 2010). And so did the others in the beginning, as a French public representative recalled:

> "when we began having this discussion with them, the narrative put forth by the platforms was: if your citizens are becoming radical it is because you aren't doing your job well (…) The fight against radicalization is the government's responsibility, so we consider that we shouldn't have to police our platforms."

But since the 2015-2017 wave of terror attacks perpetrated in Europe and the United States by the self-proclaimed Islamic State (IS, Daesh, hereafter ISIS)[6], private and public-private online CT/CVE measures have proliferated, seemingly at the initiative of Google, Facebook and to a lesser extent Twitter. How might this shift to self-regulation be explained? To apprehend it, this section first emphasizes the specificity of terrorism as a communication-dependent form of political violence, before discussing the legal and reputational pressures faced by the companies in the context of the rise of ISIS.

1.1. Terrorism as a communication-dependent form of political violence

Terrorism and the media have a longstanding relationship that has been thoroughly explored. As a form of political violence, terrorism is characterized as a weapon of the weak because it is "propaganda of the deed", that relies upon the communication of acts of violence to a larger audience to multiply their effects (Jenkins, 1974). Given that fringe groups with radical beliefs are not represented within mainstream politics and the public sphere, terrorism can be seen as an extreme way for political actors to force their way into public debate, by committing violent acts which are engineered to match newsworthiness criteria (Weimann and Winn, 1994). Brigitte Nacos observes four communicative objectives shared by all terrorists: a) public attention and intimidation through the effective hijacking of news cycles in target societies in order to spread fear; b) the recognition and publicization of their grievances and demands; c) the respect and sympathy of the constituencies on whose behalf they claim to act (propaganda); and d) a certain degree of legitimacy, obtained when high-level officials respond to terrorist messaging, thereby indirectly conferring upon them the status of a force to be reckoned with (2016).

To reach these objectives, terrorist organizations throughout history have always used the most advanced information and communication technologies of their time (Hoskins and O'Loughlin, 2010; Nacos, 2016). They were quick to develop an online presence on the WWW[7], and later to appropriate social media, taking advantage of the newest platforms as they emerged (Nacos, 2016; Schmid, 2013; Weimann, 2014). Like for other users, social media platforms offer terrorists several advantages relative to other means of communication, namely: "interactivity, reach, frequency, usability, immediacy, and permanence[8]" (Weimann, 2014: 2; Neumann, 2013). Most importantly, while terrorists were once largely reliant on the mainstream media to publicize their deeds and fulfill their four-fold communication objectives, social media allows

them to bypass such mediation and remain in control of the image they project, by producing and disseminating their own content at a potentially global scale, in what has been called "mass self-communication" (Nacos, 2016: 83). Jytte Klausen et al. date the beginning of efficient social media use for recruitment and propaganda by jihadist terrorists around 2010, when al Qaeda launched its *Inspire* e-magazine promoting "Do-It-Yourself terrorism" (2018: 9).

It is hence the unanticipated exploitation of their services by violent organizations which initially got social media companies caught up in terrorism-related issues. Today, while terrorists have notably "migrated" most of their logistics and in-group communications to encrypted services (Weimann 2014), they still strive to maintain a presence on mainstream platforms because they remain the best way for them to reach the largest possible 'captive' audience of victims, potential recruits and sympathizers. Indeed, by virtue of their very existence and scale, YouTube, Facebook and Twitter exercise an influence on terrorism-related phenomena, making them a key frontline in the struggle between terrorist organizations and the governments seeking to rein them in (Jensen et al, 2018).

1.2.Public pressures surrounding terrorist contents online

Conversely, the largest platforms are also identified by public authorities as the heart of their efforts to fight terrorists' online presence. According to one author of the French government-mandated report on countering online hate, "terrorism changed everything*", and ISIS attacks constituted a wake-up call for public authorities, renewing their interest in regulating the platform economy, within the broader context of the so-called 'techlash' (Badouard, 2020; Helberger et al, 2018). In addition to mounting threats of regulation at the EU and national levels, social media use by terrorists was publicly framed by European leaders as a pressing security issue. Facebook, YouTube and Twitter garnered specific focus as they were publicly assigned a share of the responsibility, even as they became increasingly integrated to the emerging online CT/CVE governance regime.

These developments can be traced back to the aftermath of the February 2015 *Charlie Hebdo* and Hyper Cacher attacks in France, when Interior Minister Bernard Cazeneuve made an official visit to top Silicon Valley executives from Facebook, Apple, Twitter and Google, to convince them to be more reactive to public demands. Before leaving France, he declared on national television that the largest platforms were avenues for "open access terrorism", deploring that "90% of individuals who turn to terrorism, shift through the internet*[9]" (Télé matin, 8/02/15). Such a statement drawing a direct link between the services operated by the companies he was visiting, and the traumatic events of the previous months signaled the beginning of a securitization process[10] linking terrorism and social media in the French public sphere. In effect, ensuing private negotiations resulted in the establishment of a permanent multistakeholder arrangement called the "Groupe de contact permanent" gathering Apple, Microsoft, Google, Twitter, Facebook and law enforcement around the issue of online terrorism (Tréguer, 2019).

In the UK, a similar process of securitization unfolded after the country was hit by its own wave of ISIS attacks in 2016-2017. The launch of a French-British Action Plan against terrorist use of the internet in June 2017 shows that British politicians found an ally in the newly-elected Macron administration, which made the struggle for a "safer" internet a national priority and guiding principle of French foreign policy (Tréguer, 2019). The two administrations took another trip to the Silicon Valley in January 2018 along with Germany, thereby forcefully demonstrating alignment between the three largest European powers (and markets).

The trio also used multilateral forums to set the terrorist contents online issue on the agenda, despite strong initial opposition from the United States. They were successful in high-level fora such as the United Nations (Resolution 2354; Joint Leader's Event on Preventing terrorist use of the internet), the G7 (2016 Leader's Meeting in Ise-Shima, 2017 Leaders' Summit in Taormina, 2017 and 2019 Interior Ministers' Meetings, respectively in Ischia and Paris), and the European Union. Agenda-setting efforts were especially successful in this latter arena, where they resulted in: a) the establishment of the EU Internet Forum, which was used for continuous public-private dialogue on terrorist contents online, and recognized by interviewees as instrumental to the creation of common understandings and robust working relationships at the highest levels; b) the launch of the Europol Internet Referral Unit (IRU) within the European Counter Terrorism Centre; and perhaps most importantly c) the TERREG legislative proposal.

Regulatory pressure on content issues in the EU had been growing in general, resulting in the deployment of 'soft law' instruments (Code of conduct on hate speech, Code of practice on disinformation) (Coche, 2018; Gorwa, 2019; Badouard, 2021). Meanwhile, at the national level, Member States of the aforementioned trio proposed and adopted a number of 'hard' laws on online content (NetzDG, loi Avia, Online Harms) (*ibid.*). At the prospect of a proliferation of unilateral, national initiatives threatening to fragment the European internet, it became pressing for the EU to propose its own union-wide enforceable legislation (which was also preferrable from the platforms' standpoints). Terrorist contents online were identified as the priority by member states (Seehofer and Collomb, 2018), and in this context the TERREG was drafted and debated in record time. However, even as the narrative of social media platforms as dangerous spaces to be secured took hold in the public sphere, Google, Facebook and Twitter were consistently both blamed and associated to state efforts and negotiations. This suggests the recognition by public authorities that "new actors, technologies and regulations become necessary for the state to handle the surge in public and private communications entailed by digital technologies and keep its traditional techniques of power afloat" (Tréguer, 2019:155; Arpagian in Musiani et al., 2016; Gorwa, 2019). Indeed, large platforms were extensively consulted on the TERREG, and the resulting proposal is one which, interviewees emphasized, "mostly codifies" what Google, Facebook and Twitter are already doing, except for the one-hour timeframe for the deletion of referred contents, which will require small adjustments[11].

The question of social media's potential to produce radicalization and incite violent action seems central to efforts to curb the spread of terrorist contents online. Interestingly, however, during the interviews conducted for this research, state officials rather framed their actions as a matter of countering illegal contents online (similar to the logic behind NetzDG), thereby dismissing the academic controversy around the link between the consumption of terrorist contents and violent radicalization[12]. This suggests that public authorities do not entertain false hopes as to the real-world effects of TERREG. Rather, it seems they adopt a two-faced narrative when it comes to terrorism and social media, with one strand pushed toward the public by high-level politicians after terror attacks (social media causes terrorism), and a more nuanced set of arguments advanced by specialized civil servants in their relationship to other stakeholders and experts. Therefore, it appears the public securitization of social media platforms was largely aimed at exploiting the natural vulnerabilities of these firms' trust-based business models (Culpepper and Thelen, 2020), in order to push them to the negotiation table.

Lastly, it must be noted that governments were often helped by the press in inflicting reputational damage upon the platforms on the issue of online terror. For instance, *The Times*

revealed in 2017 that some of the world's biggest companies were being advertised on extremist YouTube videos, and thereby indirectly contributed to financing groups such as ISIS and Britain First (Mostrous, 2017). A media storm ensued, followed by a boycott of YouTube and other Google services by some of their largest advertisers -including Walmart, Nestlé, Pepsi, and AT&T, causing a drop in the company's share value and an estimated ad revenue loss around $750 million (*Business Insider,* 2017). This example illustrates the business case for social media companies' involvement in CT/CVE, and how "commercial objectives become sometimes grafted onto new security roles" (de Goede, 2018: 26).

It remains unknown which of legislative pressures or reputational threat was decisive in forcing the firms to take up CT/CVE, and interviewees had differing opinions on this matter. Nevertheless, studies show that voluntary corporate measures are an efficient tool of corporate power to preempt public support for stricter government regulation, and that single-actor self-regulatory initiatives are "a way [for firms] to improve their bargaining position with other actors, to win public relations points, and to evade more costly regulation" (Abbott and Snidal, 2009:71; Gorwa, 2019; Malhotra et al., 2019). Nevertheless, a combination of these two factors[13] led to Google, Facebook and Twitter's self-regulated commitment to CT/CVE, as analyzed in the next section.

2. Social media corporations' self-regulated commitment to CT/CVE

Far from the lawless Wild West of the internet which they are often made out to be, Google, Facebook and Twitter can be considered "new governors" of online speech and democracy (Klonick, 2018). Through their extensive content moderation apparatuses, these firms set and enforce guidelines applicable to their vast user communities on prohibited contents and behaviors (Common, 2020; Gorwa et al., 2020). While these normative instruments are softer than those of states, their potentially global reach is especially relevant to the struggle against a transnational phenomenon (terrorism) taking place on a transnational medium (the internet). Google, Facebook and Twitter are therefore in a unique position to contribute to the global governance regime on terrorist communications (Ganesh and Bright, 2020; de Goede, 2018). In this sense, this section argues that their CT/CVE involvement goes beyond "privatized" enforcement (Coche, 2018: 3). Indeed, more than "a mode of security 'outsourcing' [it] involves a process of authorization and appropriation" of their new security role (de Goede, 2018: 26). The following subsections hence provide an analysis of how the studied firms have appropriated CT/CVE through self-regulation, as declined in four strands of action: policymaking, moderation, human resources and private multilateralism[14].

2.1.Private policymaking: depoliticizing terrorism?

Among scholars, the controversy around the notion of 'terrorism' is commonly admitted (Weinberg et al, 2004), and the phrase "one man's terrorist is another man's freedom fighter" has become a platitude. That said, those who are faced with the task of countering terrorism must agree upon a working definition, which may be even more complex for those who must not only define 'terrorism' but also 'terrorist content', and differentiate it from 'hate speech'. The matter of how private actors appropriate this debate is therefore a significant one, which social media corporations tend to shy away from. Their reluctance was illustrated by their lack of response when the Chairman of the US Senate hearing on online extremisms asked company representatives whether they had agreed on a shared standard for what constitutes terrorist content (C-SPAN, 17/01/18). During the interviews conducted for this research, however,

Facebook and Google representatives expressed awareness of the issue's contentiousness, even as they emphasized the objectivity and apolitical character of their own approaches.

Of the three companies studied, Facebook is the most explicit in its terrorism ban, in keeping with its reputation as the strongest ambassador of a 'safe' corner of the internet. The company explicitly prohibited credible threats and "support for violent organizations" as soon as January 2011 (Facebook, 2011). In a considerably proactive move, the company also operates its CT policy based on its own definition of "terrorist organizations and terrorists", made public under its "Dangerous Individuals and Organizations" policy, a living document which was updated six times since December 2018 (Facebook, 2021). According to the company's France representative :

> We aggregated our own definition on the basis of different definitions and expertise. Recently, it was a little bit challenged by someone from the UN who said: what about national liberation fronts who commit terrorist actions? This is complicated for us because we don't want to get into politics, it isn't our place to distinguish the good terrorists from the bad terrorists, those who commit violence in pursuit of a legitimate aim, and those who do not. So, we try to objectivize as much as possible by considering that all acts of violence aimed at killing people are against our Standards.*

In-house CT expert Brian Fishman confirmed this intent to objectivize terrorism, insisting that Facebook's was "a pretty standard academic definition of terrorism that is predicated on behavior" (Cruickshank, 2017). In this sense, Facebook's definition of terror is both a statement of proactiveness, and an expression of the company's reluctance to engage with the intrinsically political aspects of the matter.

Twitter, in line with its brand as the "free speech wing of the free speech party" (Halliday, 2012), was slower than its two competitors in integrating terrorism to its Terms of Service. In keeping with US law, the company only banned "direct, specific threats of violence against others" (Jeong, 2016) until 2015-2016, when a series of harassment scandals led to their ban on indirect threats of violence and the incitement thereof in an extended "Abusive behavior" policy, which also explicitly forbids "threatening or promoting terrorism". The company hasn't published a definition of what it considers "terrorism" but seems to employ the term interchangeably with "violent extremism". Indeed, the Safety section of the Twitter Rules includes a policy on "Violent extremist groups", complete with a definition of what constitutes a "violent extremist organization" for the company (Twitter, 2020b).

For YouTube, Community Guidelines relative to terrorism are elaborated by Google's Public Policy team. Due to public calls deploring the availability of al Qaeda and Iraq War-related materials on YouTube and Google+, the company was forced early on to adapt its policies (Neumann, 2013). In 2008, "graphic violence" and the direct incitement thereof were banned, and by 2010 a specific "promotes terrorism" button was added for users to report videos, in "one of the first instances in which [a] private compan[y] took initiative to police their own sites after being subject to public pressure" (Hughes, 2018). The term "terrorist" first appeared in the YouTube Guidelines in 2009, as an example of "content that's intended to incite violence or encourage dangerous, illegal activities that have an inherent risk of serious physical harm or death" (a phrase which replaced the term "bad stuff" in the previous version). In 2010, the company told CNN that it "remove[d] all videos and terminate[d] any account known to be

registered by a member of a designated Foreign Terrorist Organization (FTO) and used in an official capacity to further the interests of the FTO" despite this not being mentioned in their Guidelines. Still today, Google is the most cautious of the three firms with regards to taking an official stance on what it considers terrorism, violent extremism or violent criminal organizations, as the company has not publicly defined any of these terms, despite using them all.

Whether they are proactively formulating definitions of terrorism (Facebook), being ambiguous about the relationship between violent extremism and terrorism (Twitter, Google), or cautious about proposing their own definitions (Google), social media corporations' private policymaking efforts all point to the same thing: reluctance to engage with the politics inherent to the notion of terrorism, which nevertheless inevitably reappear through implementation.

2.2.Content moderation

2.2.1. From flagging to automation

Over the years, the reliance on humans (users and moderators) to take down undesirable contents came under increasing criticism for being inefficient with respect to the availability of terrorist contents (Neumann, 2013). In response, social media firms are progressively complementing the flagging system with active monitoring, as a result of their massive investments in artificial intelligence (AI) to filter their platforms (Gillespie, 2020; Gorwa et al., 2020). Today, "algorithmic moderation", understood as "systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome (e.g. removal, geoblocking, account takedown)", has become instrumental for the platforms to keep up with increased demands for responsibility from states and the public (Gorwa et al., 2020:3). Various technologies are deployed by Google, Facebook and Twitter to monitor their respective services for terrorist activity, including content "hashing"/matching to prevent previously removed content from being reuploaded, predictive/classification systems to identify and refer terrorist contents to moderators for context assessment (*ibid.*), and profiling software to identify terrorist users based on their behavior (used specifically by Twitter). According to a Facebook CT representative, due to their potentially damaging effects on free speech, such technologies are deployed only once they achieve a high success rate (a 30%-20% error rate is still too risky). They have, however, made spectacular progress in the past four years, so much so that the latest available data indicates that 99,7% of terrorist contents removed on Facebook (2020) and 98% of those on YouTube (Google, 2020) were detected before they were flagged, and 94% of Twitter accounts deleted for terrorism were detected by the company's proprietary software (2020a). In fact, automation has been instrumental in de-platforming ISIS (Berger, 2016; Conway, 2020; Ganesh and Bright, 2020).

More than a simple streamlining process, the shift from flagging to automation in order to deal with terrorist content demonstrates a change in the corporations' mindset on platform governance. This is particularly the case for Google, which was known to "remove only content that breaks laws or its terms of service, and only at the explicit request of users, governments, or courts" (Musiani et al, 2016:16; Neumann, 2013). Indeed, algorithmic moderation surpasses the demands of the law, which exempts platform firms from responsibility for content hosted by their services, only asking of them good faith in removing objectionable contents (US) or the removal of illegal contents when brought to their attention (EU). In fact, the move to automation can be seen as a further sign of the "technological solutionism" which runs in these firms' corporate DNA, as they "yet again try to solve human problems with new technology"

(Singer and Brooking, 2018: 21; Common, 2020). France, the UK and Germany actively sanctioned algorithmic moderation by encouraging it through the EU Internet Forum, to the point that the European Commission's initial proposal for the TERREG sought to impose the use of "proactive measures" similar to those deployed by Google, Facebook and Twitter, on all platforms vulnerable to terrorist exploitation. This illustrates Félix Tréguer's observation that security assemblages may be used by states to avoid their own legal constraints and commitments, in the name of the efficiency of national security (2019). Indeed, the "proactive measures" provision was scrapped by the EU Parliament, for violating the e-Commerce Directive's prohibition of "general obligation[s] (…) to monitor the information which they transmit or store, [and] to seek facts or circumstances indicating illegal activity" (Article 15).

The shift from flagging to automation also carries significant impact in terms of the discretion exercised by Google, Facebook and Twitter on which individuals and groups can be excluded from their global public spheres under CT policies. It was ISIS use of social media that prompted the political agenda-setting of the terrorist content issue (and to a lesser extent al Qaeda's). Calls to securitize social media from terrorism were exclusively made with reference to these two groups, with little concern at the time for other types of extremism (including far-right). Progress was measured against the availability of these particular groups' propaganda on the platforms, both by the firms themselves (transparency reports, blog updates) and by state authorities. Indeed, algorithmic moderation tools are deployed based on the expectations placed on the platforms by various stakeholders, which "substantially affect not only the design of the system itself, but also the ways in which that system is used to then act upon and potentially moderate content" (Gorwa et al., 2020). Proactive detection and takedown tools, then, were designed specifically for ISIS and al Qaeda propaganda (Cruickshank, 2017), and are therefore much less successful in curbing other expressions of violent extremism, be it other jihadist contents (Conway et al., 2017) or far-right ones (Berger, 2016). Because the tools developed by the platforms are trained to recognize easily identifiable characteristics, they cannot recognize "terrorism" but only specific visuals, e.g. a graphic charter or logo. That is why they were specifically well-equipped to deal with ISIS content, because the group had a "'branding strategy" that rendered its propaganda readily identifiable to platform's algorithmic moderation tools (Conway, 2020:109). For instance, YouTube's AI tools developed for ISIS contents were trained with over 2 million hand-reviewed ISIS videos, and are therefore non-transferrable to other groups (C-SPAN, 17/01/18). In particular, such methods of algorithmic moderation are ill-equipped to deal with less centralized forms of extremism, most notably far-right extremism which is a notoriously fragmented and nebulous 'scene' (Conway, 2020; Ganesh and Bright, 2020).

2.2.2. The necessary politics of terrorism

When it comes to terrorist content online, MacKenzie Common's concerns about inconsistent enforcement are clearly verified, as firms apply their blanket, vague bans on "terrorism" in an unequal manner, according to technical capacities, and "accepted narratives" about Islamism (2020: 140). The public scrutiny of just two groups, in conjunction with blanket policies, grants the firms a considerable degree of discretion in the implementation of their ban on "terrorism" and "violent extremism", as their actions against other groups and ideologies remain opaque[15]. This was made especially visible in the case of the 2019 Christchurch attack, which showed how unprepared they were to face far-right terrorism (Common, 2020). Christchurch brought to light the discrepancy between broad policies against terrorism, and their inconsistent implementation. And this discrepancy is sometimes deliberately obscured by corporate

officials. After Christchurch, Facebook announced that it was extending its hate speech policy to include white nationalism and separatism, generating a lot of good publicity. However, going back to the US Senate hearing a year before, when corporate representatives were asked what they were doing against white supremacist violence, Facebook's Monika Bickert clearly suggested that Facebook was already actively fighting white nationalist content (C-SPAN, 17/01/18), while the company's post-Christchurch communications are an admission to the contrary. Similar responses were given by Twitter and YouTube representatives, which were equally unprepared for the attack. Of course, the case of right-extremist violence is further complicated by the fact that it is generally framed by governments as 'hate' rather than 'terrorism' (Europol, 2020), which refers to a different governance regime offline and online. But this example illustrates MacKenzie Common's argument that vague content policies are beneficial to the platforms, by allowing them to be "flexible and able to react to situations quickly" (2020: 138).The politics inherent to terrorism that the companies sought to avoid through their policymaking hence inevitably resurface through the "enforcement stage" of content moderation (Common, 2020). This is especially true in the aftermath of attacks, when public scrutiny and awareness are temporarily heightened, allowing for a glimpse into otherwise opaque processes.

Google (2020) and Twitter (2020b) openly rely on government lists of terrorist organizations for the implementation of their CT policies[16]. The problem with this, however, is that designations are political, and that states tend to name *foreign, transnational* organizations as terrorist but not domestic ones[17]. In relying upon these lists, the social media giants hence reproduce the differentiated treatment afforded to various types of violent extremism that is already prevalent within counterterrorism establishments and the media (Nacos, 2016). A Google representative emphasized this as a key area for governments to take the lead, and one in which the corporation was reluctant to make its own assessments:

> If you look at the US or the UK, the UN, the EU, many of them have lists of terrorist organizations. (…) If you're a company, you can look at these lists and say: this group has (…) gone through a process that's been vetted by a democratic government, and that's one excellent authority to look to when you're trying decide what your policies should be (…) [But,] when it comes to far-right, governments have been more reluctant to name groups to such a list. (…) It gets much harder when a group isn't named. (…) And in some cases, these groups may be running for office, they may be accredited political parties, or they may be engaged in a set of actions in the real world but their videos online don't cross the same line; or they come close to the line but don't quite cross it in the real world and on the platform. [This is] an area that's very tough for companies.

In order to deal with this issue, Google now justifies that "Content produced by violent extremist groups that are not government-listed foreign terrorist organizations is often covered by (…) policies against posting hateful or violent or graphic content" (2020), while Twitter grants itself the discretion to de-platform un-designated terrorist groups (2020b). Departing from state-designated lists, however, is risky from a PR standpoint because it leaves the firms vulnerable to accusations of censorship and raises questions of legitimacy. Facebook, nonetheless, has taken this approach a step further, and maintains its own confidential list of terrorist groups, with specific rules for hybrid groups. For example, according to a Facebook representative, posts from Lebanese users who praise Hezbollah in the context of local elections

are not deleted, even though the group is listed as an FTO. This illustrates Facebook's ambition to acquire the means to respond quickly and at scale to current events. Along those lines, the company also announced in 2019 the creation of a Strategic Response Team, an in-house multidisciplinary group of experts focused on avoiding inflammatory use of their services by developing an "understanding [of] the historical, political and technological contexts of countries in conflict" (Facebook Newsroom, 2019). Likewise, having a look at Facebook's job openings within the Counterterrorism and Dangerous Organizations Team shows the firm is expanding its in-house CT expertise by hiring experts of other forms of violent extremism, such as left-wing or eco-terrorism.

All in all, the apolitical and non-specific language used in the corporate policies described in the previous subsection masks the necessarily political reality of implementing terrorism bans, leaving the companies a considerable margin of discretion.

2.3. Capacity building through human resources

The deployment of technology to deal with CT/CVE-related contents did not prevent the companies from implementing substantive organizational changes. Contrary to popular belief, automated moderation has had the opposite effect of reducing manpower, and Facebook, Google and Twitter have hired largely to adapt to their new role. All of them have internal policymaking and implementation departments, with a) public policy and/or government relations teams tasked with establishing and communicating their CT/CVE action to the world; b) teams of lawyers charged with verifying the legality of government requests or company actions in the face of challenges; and c) teams of moderators who enforce their rules. Policymaking is usually located within company headquarters in California, while other departments are spread throughout the world. As a Facebook France representative put it, there are both 'quantity' and 'quality' components to corporate CT/CVE capacity building through human resources.

The quantitative component corresponds to the large numbers of "invisible" workers who support the platforms' policy implementation (Musiani et al, 2016: 5). These are global networks of moderators spread across the largest possible number of time zones, ensuring CT requests are processed 24/7. Moderators treat requests for rules infringements including terrorism, reviewing the contents which are flagged by users, trusted flaggers (including IRUs) or AI, and deciding whether to leave them up or take them down, notably by assessing the context of posts[18]. Their ranks have grown substantially over the past years. Facebook claims to have 30,000 employees specialized in 'Safety and Security' (DLDconference, 2019), up more than 50% from 2017, including approximately 8,000 content reviewers, compared to 4,500 in late 2017 (Cruickshank, 2017). Google has reportedly more than 10,000 moderators specialized in "Safety and abuse". These teams are often contracted and/or outsourced to countries with cheap labor such as the Philippines or India, and their working conditions are notoriously far from those of the platform firms' own employees (Roberts, 2019). Leaks have revealed that moderators at Facebook, for instance, have to memorize the names and faces of over 600 terrorist leaders, and are expected to assess given pieces of content within a 10-30 second timeframe (Solon, 2017). It is noteworthy that none of the soft or hard instruments proposing to regulate moderation include provisions on the working conditions of moderators. Nevertheless, it is due in large part to them that the public representatives interviewed for this research are increasingly satisfied with the timeliness and moderation practices of the studied platforms.

Turning to the firms' more public faces, company representatives highlighted "qualitative" hiring as a key element in adjusting to their CT/CVE mission, notably through the appointment of in-house experts around 2016, to work with policymaking and engineering departments. Previously, the corporations had relied on outside consultants for this work, so their hiring in-house CT professionals reveals that another threshold has been crossed in their ownership of terrorism-related matters. Often, these new employees have academic or public service backgrounds, which facilitates the establishment of fruitful working relationships with the public sector, especially in the US (Arpagian in Musiani et al., 2016; Tréguer, 2019). Facebook is the company which has pushed this the farthest: in total, it reportedly counts more than 200 employees who work specifically on terrorism from different professional standpoints (legal experts, engineers, IT specialists, etc.). The company has hired scholars and former government CT consultants to manage its in-house CT/CVE. Brian Fishman, Facebook's Counter-terrorism Director, was previously Director of research at the Combatting Terrorism Center at WestPoint and is reportedly on ISIS' kill list; while regional Heads of CT policy also have strong academic and government backgrounds. Similarly, Google's Global Policy Lead for Counterterrorism is Dr. William McCants, a renowned specialist of militant Islam, who consulted for the US State Department's Office of the Coordinator for Counterterrorism and formerly worked at the Brookings Institution. Google's French representative also emphasized that the company had hired free speech legal experts, in order to "push back" against abusive demands from the authorities. In contrast to Facebook and Google, Twitter has not hired in-house terrorism experts, although the company has reportedly hired former law enforcement officials who, in relation with outside consultants, oversee the platform's CT/CVE action (C-SPAN, 17/01/18). Indeed, while Google and Facebook both emphasize the continued need for humans to complement technology because of its fallibility, Twitter is more upfront about delegating terrorism-related decisions to its software and has not communicated an estimate number of content moderators in its employ.

Finally, in addition to their in-house capacity building, all studied companies consult external entities specialized in terrorism to stay on top of developing trends. These include start-ups (Moonshot CVE, SITE Intelligence Group), think-tanks (Institute for Strategic Dialogue, Institut français des Relations internationals), and public bodies such as Europol's IRU, who also have "trusted flagger" status, allowing for prioritization of their takedown requests.

2.4. The private multilateralism of the GIFCT

Social media corporations have also displayed pro-activity in CT/CVE by joining forces. Their collaborative effort ultimately lead to the launch of the Global Internet Forum to Counter Terrorism (GIFCT) in May 2017 by YouTube, Facebook, Twitter and Microsoft. This private multilateralism initiative is commonly justified by the phenomenon of terrorist migration (Weimann, 2014), which requires industry-level responses, in addition to facilitating negotiations with public authorities by providing a 'one-stop shop' to discuss online terrorism with 'the private sector'. The Forum is governed by a committee of senior representatives from the founding firms. Other companies can apply for membership provided they satisfy six criteria set out by the founding members, who retain the discretion to admit or reject applicants. In addition to these criteria, candidates must also "agree that governments will not be able to remove terrorist content directly from company platforms" (Facebook, n.d.). Decision-making power within the GIFCT thus remains in the hands of the founding members, who laid down

the rules which newcomers must conform to on what constitutes a commitment to "significantly disrupt[ing] terrorist exploitation of the Internet". In this sense, the GIFCT can be seen to allow its founding members to further extend their influence within the industry.

The GIFCT has three operational "pillars", the most important being the "technology pillar" which comprises the "shared industry hash database" (SIHD) for terrorist contents. Launched in early 2017, the SIHD provides for the attribution of "a string of data meant to uniquely identify the underlying content", or hash, to terrorist contents taken down by member platforms (Gorwa et al., 2020:4). Hashes are added to the shared database, thereby allowing other participating companies to scan their own services for the content. As of July 2020, 13 companies shared access to the database, which contained 300,000 unique hashes (Google, 2020). A number of questions on the database remain unclear, such as whether being a GIFCT member is a prerequisite for access, or how much of the process is automated. Little is known about the precise contents of the database. Initially, the SIHD was only used for ISIS and al-Qaeda-related content, specifically the most egregious violent content which is banned on all member platforms (Cruickshank, 2017); however, following the Christchurch attack it has been extended to other violent events (Gorwa et al. 2020). Furthermore, the companies have different policies on how the database fits into their enforcement mechanisms. For instance, Microsoft reportedly only adds to the database, but does not scan its own services for further takedowns, while YouTube is scanned for contents available in the database, but not other Google services (Brandom, 2016). Company representatives strongly emphasized that the GIFCT was not meant to harmonize content policies, as each member retained their own takedown thresholds.

The second pillar, "knowledge sharing and information", encompasses GIFCT efforts to spread content moderation best practices with smaller organizations lacking the resources or manpower to develop their own solutions to terrorist exploitation. Getting smaller social media platforms to join the GIFCT was identified as a key goal by both corporate and state representatives in interviews conducted for this research. In practice, this work is conducted through Tech against Terrorism, a public-private partnership funded by the GIFCT companies and the governments of Spain, Switzerland and South Korea.

Finally, the GIFCT's third "research" pillar was launched in February 2019. It consists in financing the Global Research Network on Terrorism and Technology, a network of 8 partner institutions led by the Royal United Services Institute, to conduct policy-oriented studies on terrorist uses of the internet.

Conclusion

Drawing on publicly available sources and interviews, this article has provided an empirically-informed overview of Google, Facebook and Twitter's engagement with CT/CVE, highlighting four main dimensions of their involvement, namely policymaking, content moderation, human resources and private multilateralism. It finds that the studied corporations are moving from a reactive posture, in the context of reputational threats and the prospect of hard regulation in the EU, to an increasingly proactive one as global security actors in CT/CVE. They display both a commitment to self-regulation and creativity in their engagement with this policy area, surpassing what is legally required of them in the EU and US (for now). Effectively, these private actors are playing a key role in the development and implementation of the nascent online CT/CVE regime. Somewhat similar to terrorism financing which is more codified, this

is indicative of a new logic of governance in the policy areas of online CT/CVE, where platform firms have become a 'link' in the "chain of security" (de Goede, 2018) by virtue of their self-regulation, and in the logic of "global security assemblages" (Abrahamsen and Williams, 2010; Tréguer, 2019). Whether these findings are applicable to other tech firms, especially non-Western ones and in non-Western contexts, is an interesting question for future research, as indicated by TikTok's failed attempt (so far) to join the GIFCT (Birnbaum, 2019).

The current terrorist content governance regime emerged from the chaos, urgency and emotion of ISIS attacks, leaving little room for public debate on how efficient, or even desirable, it is to involve social media corporations in CT/CVE in the first place. As argued by critics, the overreliance on algorithmic moderation in particular, as accelerated by Covid-19, presents risks to freedom of expression by essentializing political choices over categories of 'problematic speech', embedding bias through inconsistent enforcement, whilst covering it in a veneer of objectivity through technology (Common, 2020; Gillespie, 2020; Gorwa et al., 2020). As shown in this paper, the resulting state of things is that social media firms are left with a considerable margin of discretion to decide which political groups to silence and in what measure. The same veil of secrecy surrounding content moderation in general (*ibid.*), characterized by vague rules, opacity and inconsistent enforcement (Common, 2020) also plagues terrorist content issues. In fact, these problems may be heightened in the case of terrorism, where secrecy is also justified by the need to keep terrorists in the dark about ways to bypass moderation systems. As CT policy and CVE initiatives are famously prone to civil liberties restrictions and discrimination, questions of accountability are key avenues for further research, while transparency appears a necessary first step.

As shown by repeated moderation scandals, it is in the firms' long term interests to communicate more on the complexity of their work and the challenges they face – if only to reverse the narrative that they are passive in the face of extremism, hate and foreign interference. And albeit to varying extents, it seems the studied firms are slowly opening up to independent input from governments, the public and researchers on content issues. One promising development is the launch Facebook's Oversight Board, which will no doubt be brought to assess terrorist content cases in the near future. Mark Zuckerberg has repeatedly called for more government regulation, and Twitter is multiplying user surveys about content policy following its Donald Trump ban. This is good news for those seeking to open up the 'black box' of corporate platform moderation, because the sources available to study are multiplying, including Oversight Board precedents, extended transparency reports, Terms of Service archives, job postings, corporate blogs, parliamentary proceedings and leaks.

Notes

[1] Countering violent extremism (CVE) approaches imposed themselves following the failure of "hard" counter-terrorism (CT) strategies post-9/11, which usually involve the use of force and are aimed at disrupting terrorism through intelligence, military action and law enforcement. "Soft" approaches, by contrast, aim to act upon the hearts and minds of individuals who are vulnerable to violent extremism. They include the prevention of radicalization, counter-radicalization and de-radicalization efforts. CVE (or PVE, Preventing Violent Extremism) is a related concept and subset of "soft" CT which refers to initiatives targeting the drivers of radicalization (Bjola and Pamment 2019; Frazer and Nünlist, 2015). CVE is broader in scope than CT in that it addresses radicalization into politically-motivated violence, and not just into designated terrorist organizations. On a side note, this article refers to "terrorism and violent extremisms" in an effort to emphasize the political nature of terrorist designations and account for the many groups and individuals who, without this label, also espouse political violence. The

implications of this divide between designated and un-designated groups for the subject-matter at hand are discussed in section 2.1 and 2.2 of this article.

[2] Facebook, Google (YouTube) and Twitter are at the center of this development due to their structural prominence, with user counts of respectively 2,7, 2 billion and 330 million. They are, along with Microsoft, founding members of the Global Internet Forum to Counter Terrorism (GIFCT, see section 2.4), and in relation to newer social media platforms like Snapchat or TikTok (whose user counts exceed Twitter's), they stand out for their resilience and longevity (15+ years).

[3] See for instance the leaked joint letter of French and German Interior Ministers to the European Commission (Seehofer and Collomb, 2018).

[4] Collected from the following URLs : https://blog.google/ ; https://europe.googleblog.com ; https://googleblog.blogspot.com ; https://youtube-uk.googleblog.com ; https://youtube.googleblog.com/ ; https://newsroom.fb.com/ and https://blog.twitter.com

[5] 14 interviews were conducted with representatives from Facebook, Google, the French Ministries of the Interior and of Europe and Foreign Relations, the German Ministry of Interior, Tech against Terrorism, Moonshot CVE, and la Quadrature du Net. Analysis of Twitter is based solely on publicly-available information, as the firm never responded to interview requests.

[6] Between 2015 and 2018, ISIS claimed 12 terror attacks of five deaths or more in France, Germany, the UK and the USA. The lethality of this particular group, the unprecedented scale and reach of its social media efforts and professional, centralized propaganda machine explain its centrality in the developments analyzed, (Berger, 2016; Nacos, 2016; Conway, 2020) as it was made an absolute policy priority by public authorities in their engagement the platform firms.

[7] Zelin reports that the first jihadist website ever to be launched on the Web was the Islamic Media Center's in 1991 (2013:5). The white supremacist forum Stormfront, launched in 1995, is commonly referred to as the first right-extremist website (Conway, 2017).

[8] Permanence means that due to their decentralized nature and mainstream appeal, social media platforms cannot be shut down or attacked like terrorist forums or websites were.

[9] Quotes followed by an * have been translated from French by the author.

[10] Building upon the observation that security is a social construct, securitization originates in the constructivist school of International Relations, and was first developed by Barry Buzan, Ole Waever and Jaap de Wilde (1998). It can be defined as the process whereby an issue is removed from ordinary politics, and recast as a security threat, thereby warranting the use of extraordinary means to eradicate it (*ibid.).* While the Copenhagen School's original framework postulates the performativity of language, the 'sociological' model places a greater emphasis on context (Balzacq, 2011).

[11] This is the basis of a common criticism of this legislation by online rights advocates (EDRi, La Quadrature du Net, EFF), who emphasize that only the largest platforms will be able to comply and that it will lead to over-blocking by smaller actors fearing penalties.

[12] The sequence of how on- and offline factors interact to produce a radicalized, violent individual is highly contested (Conway, 2017), and "online radicalization" is commonly perceived as an exception rather than the rule (Schmid, 2013; Klausen et al. 2018). However, it is a fair assumption that social media platforms act as facilitators and accelerators of radicalization (*ibid.*; Nacos, 2016; Neumann, 2013).

[13] To these prosaic reasons, it must also be added that an obvious element of response to the question of their involvement in CT/CVE is that Google, Facebook and Twitter's employees are affected by terrorism's adverse effects as citizens of the societies to which they belong and genuinely attempt to do their part in mitigating them.

[14] Corporate social responsibility initiatives, including the work done by Google's Jigsaw think-tank, programs such as Facebook's now-defunct Peer2Peer or Twitter's #TwitterForGood also contribute to their CT/CVE activities but fall outside the scope of content moderation.

[15] Facebook announced recently that it would start reporting on other groups.

[16] All US companies are legally prohibited from hosting content from organizations listed as Foreign Terrorist Organizations (FTOs) by the US State Department.

[17] The question of un-listed violent organizations is likely to change in the near future given the January 2021 Capitol attack and the Biden administration's subsequent joining of the Christchurch Call.

[18] Romain Badouard notes that moderation decisions have grown more complex in recent years: in addition to the binary "leave up" or "take down", a range of options in between now include downgrading contents, removing them from search results and/or recommendation systems, demonetizing them, etc. (2020).

References

Abbott, KW, Snidal, D (2009) The Governance Triangle: Regulatory Standards Institutions and the Shadow of the State. In: Mattli W and Woods N (eds) *The Politics of Global Regulation*. Princeton: Princeton University Press, pp. 44–88.

Abrahamsen, R, Leander, A (eds) (2016) *Routledge Handbook of Private Security Studies*. London ; New York: Routledge, Taylor & Francis Group.

Abrahamsen, R, Williams, MC (2010) *Security Beyond the State: Private Security in International Politics*. Cambridge: Cambridge University Press.

Badouard, R (2020) *Les Nouvelles Lois du web: Modération et censure*. La République des idées. Paris: Seuil.

Badouard, R (2021) Ce que peut l'État face aux plateformes. *Pouvoirs* N° 177(2) : 49–58.

Balzacq, T (ed.) (2011) *Securitization Theory: How Security Problems Emerge and Dissolve*. PRIO new security studies. Milton Park, Abingdon, Oxon ; New York: Routledge.

Berger, JM (2016) *Nazis vs. ISIS on Twitter:* Washington, D.C.: The Centre for Extremism at George Washington University.

Beutel, A, Weinberger, P (2016) *Public-Private Partnerships to Counter Violent Extremism: Field Principles for Action*. Final Report to the U.S. Department of State. College Park, Maryland: START.

Birnbaum E (2019) TikTok seeks to join tech fight against online terrorism. *The Hill,* 4 November. Available at: https://thehill.com/policy/technology/468884-tiktok-seeks-to-join-tech-fight-against-online-terrorism (accessed 5 July 2021).

Bjola, C, Pamment, J (eds) (2019) Countering Online Propaganda and Extremism: The Dark Side of Digital Diplomacy. Routledge New Diplomacy Studies. London ; New York: Routledge.

Brandom, R (2016) The political fight behind Facebook and Google's new terrorist content database. *The Verge*, 8 December. Available at: https://www.theverge.com/2016/12/8/13886988/terrorist-content-database-facebook-google-youtube-microsoft-twitter (accessed 8 January 2021).

Business Insider (2017) Analyst predicts the YouTube advertiser boycott will cost Google $750 million. 27 March. Available at: https://www.businessinsider.in/analyst-predicts-the-youtube-advertiser-boycott-will-cost-google-750-million/articleshow/57855322.cms (accessed 12 May 2021).

Buzan, B, Wæver, O, de Wilde, J (1998) *Security: A New Framework for Analysis*. Colorado: Lynne Rienner.

CNN (2010) YouTube defends policies in wake of complaint over al-Awlaki videos. 25 October. Available at: http://www.cnn.com/2010/POLITICS/10/25/youtube.al.awlaki/index.html (accessed 8 January 2021).

Coche, E (2018) Privatised enforcement and the right to freedom of expression in a world confronted with terrorism propaganda online. *Internet Policy Review* 7(4). DOI: 10.14763/2018.4.1382.

Common, MF (2020) Fear the Reaper: how content moderation rules are enforced on social media. *International Review of Law, Computers & Technology* 34(2). Routledge: 126–152. DOI: 10.1080/13600869.2020.1733762.

Conway, M (2017) Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research. *Studies in Conflict & Terrorism* 40(1): 77–98.

Conway, M, Khawaja, M, Lakhani, S, et al. (2017) *Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts*. Policy Report. VOX-Pol.

Conway, M (2020) Routing the Extreme Right - Challenges for Social Media Platforms. *The RUSI Journal* 165(1). Routledge: 108–113. DOI: 10.1080/03071847.2020.1727157.

Cruickshank, P (2017) A View from the CT Foxhole: An Interview with Brian Fishman, Counterterrorism Policy Manager, Facebook. *CTC Sentinel* 10(8): 8–12.

C-SPAN (2018) United States Senate Commerce, Science and Transportation Committee hearing on Extremist Propaganda and Social Media. 17 January. Available at: https://www.c-span.org/video/?439849-1/facebook-twitter-youtube-officials-testify-combating-extremism (accessed 8 January 2021).

Culpepper, PD, Thelen, K (2020) Are We All Amazon Primed? Consumers and the Politics of Platform Power. *Comparative Political Studies* 53(2): 288–318. DOI: 10.1177/0010414019852687.

Europol (2020) *European Union Terrorism Situation and Trend Report 2020*. (TE-SAT). The Hague: European Union Agency for Law Enforcement Cooperation.

Facebook (n.d.) Global Internet Forum to Counter Terrorism (GIFCT). Available at: https://counterspeech.fb.com/en/initiatives/global-internet-forum-to-counter-terrorism-gifct/ (accessed 5 July 2021).

Facebook (2011) Facebook Community Standards. Available at: Retrieved from: https://web.archive.org/web/20110127224041/https://www.facebook.com/communitystandards/ (accessed 8 January 2021).

Facebook (2021) Community Standards Recent Updates, Dangerous Individuals and Organizations (September 2019). Available at: https://www.facebook.com/communitystandards/recentupdates/dangerous_individuals_organizations (accessed 12 January 2021).

Facebook Newsroom (2019) Understanding Social Media and Conflict. Available at: https://about.fb.com/news/2019/06/social-media-and-conflict/ (accessed 5 July 2021).

Frazer, O, Nünlist, C (2015) *The Concept of Countering Violent Extremism*. 183, CSS Analyses in Security Policy. Zürich: Center for Security Studies, ETH Zürich.

Ganesh, B, Bright, J (2020) Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation. *Policy & Internet* 12(1): 6–19. DOI: 10.1002/poi3.236.

Gillespie, T (2010) The politics of 'platforms'. *New Media & Society* 12(3): 347–364. DOI: 10.1177/1461444809342738.

Gillespie, T (2020) Content moderation, AI, and the question of scale. *Big Data & Society* 7(2). DOI: 10.1177/2053951720943234.

Goede, M de (2018) The chain of security. *Review of International Studies* 44(1). Cambridge University Press: 24–42. DOI: 10.1017/S0260210517000353.

Google (2020) YouTube Community Guidelines enforcement – Violent Extremism. Available at: https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en (accessed 12 January 2021).

Gorwa, R (2019) The platform governance triangle: conceptualising the informal regulation of online content. *Internet Policy Review* 8(2). DOI: 10.14763/2019.2.1407.

Gorwa, R, Binns, R, Katzenbach, C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7(1). DOI: 10.1177/2053951719897945.

Halliday, J (2012) Twitter's Tony Wang: 'We are the free speech wing of the free speech party'. *The Guardian*, 22 March. Available at: http://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech (accessed 8 January 2021).

Helberger, N, Pierson, J, Poell, T (2018) Governing online platforms: From contested to cooperative responsibility. *The Information Society* 34(1): 1–14. DOI: 10.1080/01972243.2017.1391913.

Hoffman, J (2013) Theorizing power in transition studies: the role of creativity and novel practices in structural change. *Policy Sciences* 46(3): 257–275.

Hoskins, A, O'Loughlin, B (2010) *War and Media: The Emergence of Diffused War*. Cambridge: Polity.

Hughes, S (2018) Whose Responsibility Is It to Confront Terrorism Online? In: *Lawfare*. Available at: https://www.lawfareblog.com/whose-responsibility-it-confront-terrorism-online (accessed 8 January 2021).

Jenkins, BM (1974) *International Terrorism: A New Kind of Warfare*. Santa Monica, CA: RAND Corporation.

Jensen, M, James, P, LaFree, G, et al. (2018) *The Use of Social Media by United States Extremists*. College Park, Maryland: START.

Jeong, S (2016) The History of Twitter's Rules. *Vice Motherboard*, 14 January. Available at: https://www.vice.com/en/article/z43xw3/the-history-of-twitters-rules (accessed 8 January 2021).

Klausen, J (2015) Tweeting the Jihad : Social Media Networks of Western Foreign Fighters in Syria and Iraq. *Studies in Conflict & Terrorism* 38(1): 1–22. DOI: 10.1080/1057610X.2014.974948.

Klausen, J, Libretti, R, Hung, BWK, et al. (2018) Radicalization Trajectories: An Evidence-Based Computational Approach to Dynamic Risk Assessment of "Homegrown" Jihadists. *Studies in Conflict & Terrorism*: 1–28. DOI: 10.1080/1057610X.2018.1492819.

Klonick, K (2018) The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* (131): 1598–1670.

Krasner, SD (1982) Structural Causes and Regime Consequences: Regimes as Intervening Variables. *International Organization* 36(2): 185–205.

Malhotra, N, Monin, B, Tomz, M (2019) Does Private Regulation Preempt Public Regulation? *American Political Science Review* 113(1): 19–37. DOI: 10.1017/S0003055418000679.

Mostrous, A (2017) Big brands fund terror through online adverts. *The Times*, 9 January. Available at: https://www.thetimes.co.uk/article/big-brands-fund-terror-knnxfgb98 (accessed 12 May 2021).

Musiani, F, Cogburn, DL, DeNardis, L, et al. (2016) *The Turn to Infrastructure in Internet Governance*. London ; New York: Palgrave Macmillan.

Nacos, B (2016) *Mass-Mediated Terrorism: Mainstream and Digital Media in Terrorism and Counterterrorism*. Third edition. Lanham: Rowman & Littlefield.

Neumann, PR (2013) Options and Strategies for Countering Online Radicalization in the United States. *Studies in Conflict & Terrorism* 36(6): 431–459. DOI: 10.1080/1057610X.2013.784568.

Radicalisation: the counter-narrative and identifying the tipping point (2016) HC135, 25 August. London: UK House of Commons Home Affairs Committee.

Ritzmann, A (2017) The Role of Propaganda in Violent Extremism and how to Counter It. In: *Violent Extremism in the Euro-Mediterranean Region*. EuroMed Survey of Experts and Actors 8. Barcelona: Institut Europeu de la Mediterrania (IEMed), pp. 26–32.

Roberts, ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.

DLDconference (2019) *What Kind Of Internet Do We Want? (Sheryl Sandberg, Facebook) | DLD 19*. DLD Conference. Available at: https://www.youtube.com/watch?v=BbMo6nvlpsE (accessed 29 January 2021).

Schmid, A (2013) Radicalisation, De-Radicalisation, Counter-Radicalisation: A Conceptual Discussion and Literature Review. *The International Centre for Counter-Terrorism – The Hague* 4(2). DOI: 10.19165/2013.1.02.

Security Council resolution 2354 (2017) [on implementation of the Comprehensive International Framework to Counter Terrorist Narratives] (2017) S/RES/2354(2017), 24 May. New York: United Nations Security Council.

Seehofer, H, Collomb, G (2018) [joint letter to the European Commission by the French and German Interior Ministers, leaked]. Available at: https://edri.org/our-work/leak-france-germany-demand-more-censorship-from-internet-companies/ (accessed 16 May 2021).

Singer, PW, Brooking, ET (2018) *Likewar: The Weaponization of Social Media*. Boston: Houghton Mifflin Harcourt.

Solon, O (2017) To censor or sanction extreme content? Either way, Facebook can't win. Available at: http://www.theguardian.com/news/2017/may/22/facebook-moderator-guidelines-extreme-content-analysis (accessed 17 May 2021).

Télé Matin (2015) Les 4 vérités - Bernard Cazeneuve. 8 February. Available at: Retrieved from: https://www.youtube.com/watch?v=jlcsCHV-vKc (accessed 8 January 2021).

Tréguer, F (2019) Seeing like Big Tech: security assemblages, technology, and the future of state bureaucracy. In: Bigo D, Isin E, and Ruppert E (eds) *Data Politics: Worlds, Subjects, Rights*. Routledge Studies in International Political Sociology. London: Routledge.

Twitter (2020a) Rules Enforcement. Available at: https://transparency.twitter.com/en/reports/rules-enforcement.html (accessed 12 January 2021).

Twitter (2020) Violent organizations policy. Available at: https://help.twitter.com/en/rules-and-policies/violent-groups (accessed 12 January 2021).

Weimann, G (2014) *New Terrorism and New Media*. Research series 2. Washington, D.C.: Commons Lab of the Woodrow Wilson International Center for Scholars.

Weimann, G, Winn, C (1994) *The Theater of Terror: Mass Media and International Terrorism*. New York: Longman.

Weinberg, L, Pedahzur, A, Hirsch-Hoefler, S (2004) The Challenges of Conceptualizing Terrorism. *Terrorism and Political Violence* 16(4): 777–794. DOI: 10.1080/095465590899768.

YouTube (2009) YouTube Community Guidelines. Available at: https://web.archive.org/web/20090403124358/http://www.youtube.com/t/community_guidelines (accessed 8 January 2021).

Zelin, AY (2013) *The State of Global Jihad Online: A Qualitative, Quantitative, and Cross-Lingual Analysis*. Washington, D.C.: New America Foundation.