NEGOTIATING RESPONSES TO ONLINE TERRORISM THREATS IN THE EU

STATE-PLATFORM DIPLOMACY 2015–2019

Marguerite Borelli

NEGOTIATING RESPONSES TO ONLINE TERRORISM THREATS IN THE EU

STATE-PLATFORM DIPLOMACY 2015–2019

About the author

Dr. Marguerite Borelli is Postdoctoral Researcher at the Centre for Internet and Society (CIS CNRS) in Paris, where she works on the international Open Research Area (ORA) project "(Re-)claiming digital sovereignty in discourse, policy and practice" (ClaimSov). She obtained her PhD in 2025 from the Analysis and Interdisciplinary Research Centre for Media (CARISM), Paris-Panthéon-Assas University, with a thesis focused on the role of large social media corporations in the governance of terror-related threats online, and the public-private relations surrounding terrorist uses of the internet. Her research has been published in English and French language journals, such as *New Media & Society* and *Mots. Les langages du politique*.

Acknowledgements

This piece has undergone a long evolution before reaching its current form, and I am deeply grateful to those who have helped shape it. I would first like to thank Stuart Macdonald and Maura Conway for the opportunity to publish this research as a VOX-Pol report. The VOX-Pol team and network have been a pleasure to work with, and I sincerely appreciate the thoughtful feedback provided by the two anonymous reviewers. My gratitude also goes to Veronica Kelly for her meticulous editing and to Louise Laing for ensuring a smooth publication process. On the other side of the Channel, I am especially indebted to my PhD supervisors, Cécile Méadel and Romain Badouard, for their patient guidance and support. I would also like to thank Anne Bellon, Clément Mabi, and Gilles Jeannot for their feedback on an early version of this work, presented at the 2022 AFSP Congress in Lille. Last but certainly not least, my deepest appreciation goes to the professionals who generously agreed to take part in this research by sharing their experiences and perspectives during interviews.

Funding details

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Disclosure statement

The author reports that there are no competing interests to declare.

ISBN: 978-1-911669-86-9

Designed and typeset by Soapbox, www.designbysoapbox.com

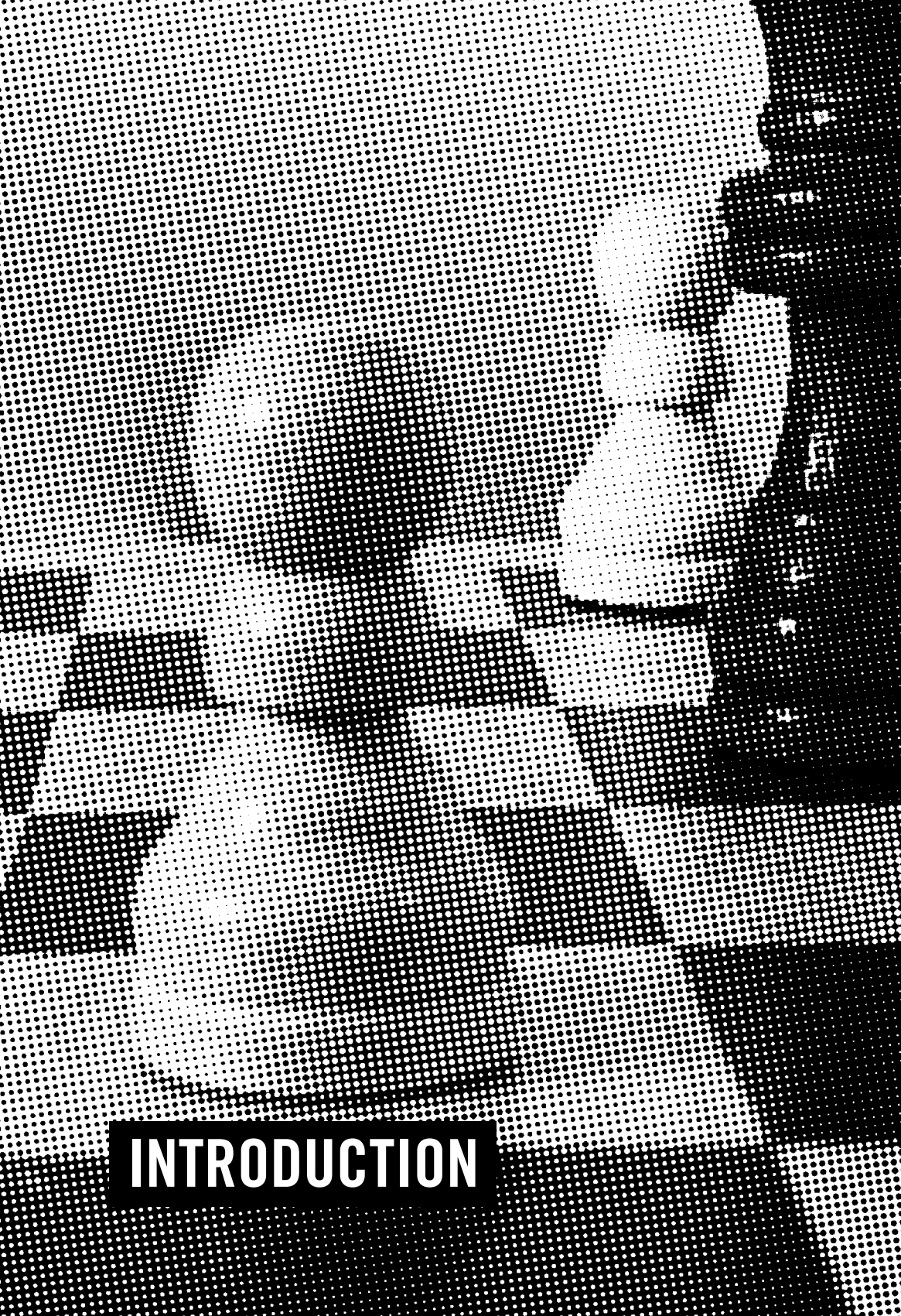
TABLE OF CONTENTS

ABOUT

THIS REPORT ANALYSES public-private relations between the European (in particular, French) authorities and Facebook, Google (YouTube) and Twitter on the subject of terrorist usage of social media between 2015 and 2019. In a qualitative approach, interview material is mobilised to investigate lived experiences of this early cooperation, focusing on how stakeholders defined and negotiated their involvement, navigated power relations and pursued strategies to establish working relations without abandoning their respective preferences. The report finds that, although cultural fault lines influenced the stakeholders' perceptions of roles, enthusiasm for 'automated' moderation was shared across the public-private boundary, justifying the development of what can be called state-platform diplomacy.

LIST OF ABBREVIATIONS

AI: artificial intelligence
CSAM: child sexual abuse material
CT: counter-terrorism
CVE: countering violent extremism
CTED: Counter-Terrorism Executive Directorate (UN)
DSA: Digital Services Act (EU)
EU: European Union
EUIF: European Union Internet Forum
Europol: European Union Agency for Law Enforcement Cooperation (formerly European Police Office)
GIFCT: Global Internet Forum to Counter Terrorism
IRU: Internet Referral Unit
IS, ISIS or Daesh: Islamic State (of Iraq and Syria)
MEP: Member of the European Parliament
MFA: Ministry of Foreign Affairs
MP: Member of Parliament
SIHD: Shared Industry Hash Database
TERREG/TCO Regulation: Regulation on Combating the Dissemination of Terrorist Content Online (EU)
TVEC: terrorist and violent extremist content
UK: United Kingdom
UN: United Nations
US/USA: United States of America



INTRODUCTION

> *Terrorism for a long time, or counter-terrorism I should say, has really been a driver on a lot of both platform policy development and regulatory policy development. I think that's going to continue to be true forever, probably. Terrorism is one of those issues where it's very emotionally charged, understandably and necessarily. That means it is quite central to the way that platforms end up talking through some of their hardest cases, and it's obviously a huge motivating topic for governments, for regulators.*[1]

IN AN EPISODE of the Tech against Terrorism podcast, Charlotte Willner, a veteran of Facebook's early Trust and Safety team and founder of the Trust and Safety Professional Association (TSPA), pointed to a puzzle. In the fraught landscape of content regulation, the past ten years have seen the emergence of a public-private, potentially global regime aimed at countering terrorism online, which stands out as an exception. Empirical research on the cooperation between the tech sector and counter-terrorism law enforcement in five countries has found that respondents on both sides of the public-private boundary agree that "significant progress has been made" in this respect.[2] Platform

1 Willner, C. and A. Craanen. 30 March 2022. Tech Policy Evolution & The Human Side of Moderating Terrorist Content (PART 1). Tech against Terrorism Podcast. https://podcast.techagainstterrorism.org/1684819/episodes/10338244-tech-policy-evolution-the-human-side-of-moderating-terrorist-content-part-1. Similar observations were made in Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press, p. 37.

2 Macdonald, S. and A. Staniforth. 2023. *Tackling Online Terrorist Content Together: Cooperation between Counterterrorism Law Enforcement and Technology Companies* (p. 36). Global Network on Extremism and Technology (GNET). https://doi.org/10.18742/PUB01–110, p. 14.

representatives[3] have attended G7 and UN Security Council meetings, engaging with political leaders on this issue at the highest levels of multilateralism and security governance. They have pooled resources and intel through the Global Internet Forum to Counter Terrorism (GIFCT). They routinely engage with various law enforcement agencies and government officials through new, tailor-made fora such as the Christchurch Call and the European Union's Internet Forum.

On the one hand, the involvement of private actors in security governance is not unprecedented,[4] particularly in cyberspace, where it is commonplace.[5] On the other hand, as highlighted by Bharath Ganesh and Jonathan Bright, countering online threats requires whole-of-government, if not whole-of-society approaches.[6] However, as Marieke de Goede points out, unlike private security companies or cybersecurity firms, social media corporations have historically been reluctant to get involved in security-related policy areas,[7] using their legal status as neutral intermediaries as a defence allowing them to get out of policing their services. So, how did we get here?

3 This report uses T. Gillespie's understanding of platforms as "online sites and services that: a) host, organize, and circulate users' shared content or social interactions for them, b) without having produced or commissioned (the bulk of) that content, c) built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising and profit" (p. 18), to which he later adds "platforms do, and must, moderate the content and activity of users, using some logistics of detection, review, and enforcement" (p. 21). Gillespie. 2018. *Custodians of the Internet.*

4 See for instance Abrahamsen, R. and A. Leander, eds. 2016. *Routledge Handbook of Private Security Studies.* London; New York: Routledge, Taylor & Francis Group. Abrahamsen, R. and M.C. Williams. 2010. *Security Beyond the State: Private Security in International Politics.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511974441.

5 Dunn Cavelty, M. 2016. "Cyber-Security and Private Actors." In *Routledge Handbook of Private Security Studies*, edited by R. Abrahamsen and A. Leander. London; New York: Routledge, Taylor & Francis Group, pp. 89–99.

6 Ganesh, B. and J. Bright. 2020. "Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation." *Policy & Internet* 12, No. 1, pp. 6–19. https://doi.org/10.1002/poi3.236.

7 de Goede, M. 2018. "The Chain of Security." *Review of International Studies* 44, No. 1, pp. 24–42. https://doi.org/10.1017/S0260210517000353, p. 26.

This report investigates early lived experiences of public-private cooperation on the issue of terror-related threats online, looking in particular at the relationship between the large US-based social media corporations Facebook, Google (YouTube) and Twitter[8] and European authorities, in particular French ones. It focuses on the period between January 2015 (*Charlie Hebdo* and Hyper Cacher attacks) and March 2019 (Christchurch attack). This timeframe can be considered that of the emergence of the European governance regime on terrorist uses of the Internet, encompassing both public authorities' efforts to place the social media-terrorism link on the agenda in France and within the EU, and the first stages of negotiations around the European Regulation on Combating the Dissemination of Terrorist Content Online (2021/784, also known as the TERREG or TCO Regulation). During this time, in an effort to pre-empt potentially costly regulation, companies implemented a self-regulatory approach to countering terrorism on their services by adapting their content rules, deploying human and technical resources to moderate their platforms better, hiring subject-matter experts, and forming the GIFCT.[9] The period also covers an evolution in the threat landscape, extending from the heyday of the self-proclaimed Islamic State (ISIS or Daesh) – marked by a series of large-scale attacks on Western

8 The report refers to Facebook rather than Meta, and Twitter rather than X, because its scope is limited to a time before their respective rebrandings. They were chosen as case studies, along with Google (YouTube), because they are the founding social media companies of the Global Internet Forum to Counter Terrorism (GIFCT), which they launched in 2017 with Microsoft.

9 Borelli, M. 2021. "Social Media Corporations as Actors of Counter-Terrorism." *New Media & Society* 25, No. 11, pp. 2877–2897. https://doi.org/10.1177/14614448211035121. Citron, D.K. 2018. "Extremist Speech, Compelled Conformity, and Censorship Creep." *Notre Dame Law Review* 93, No. 3, pp. 1035–1072; Malhotra, N., B. Monin and M. Tomz. 2019. "Does Private Regulation Preempt Public Regulation?" *American Political Science Review* 113, No. 1, pp. 19–37. https://doi.org/10.1017/S0003055418000679.

countries, and an unprecedented online presence[10] – to the loss of its territorial base in Iraq and Syria, which led to a quantitative and qualitative drop in its propaganda.[11] The relative decline in Daesh propaganda – resulting from the group's loss of capacity, together with active efforts to de-platform it – coincides with the parallel rise of a more diffuse threat of far-right violent extremism which culminated in the Christchurch attack on 15 March 2019.[12] As concern about far-right-inspired violence soared, the scope of the nascent governance regime to counter terrorism online expanded beyond the EU (Christchurch Call), and beyond ISIS, to focus more generally on "terrorist *and violent extremist* content" (TVEC) – terminology that conveniently circumvents politically charged debates on what qualifies as "terrorism". The Christchurch attack and its aftermath also constituted a milestone in the evolution of public-private relationships surrounding TVEC.[13] Symbolic of these developments are the Christchurch Call, launched by New Zealand and France in 2019, and eventually joined by the United States following the Capitol assault in January 2021, and the GIFCT's transition from an inter-industry forum to an "independent" body with its own full-time staff. This report aims to contribute to our understanding of the current TVEC governance ecosystem by going back to the initial negotiations at the EU level that made it possible, ultimately laying the groundwork for more ambitious regulation in the form of the Digital Services Act.

10 Conway, M., M. Khawaja, S. Lakhani, J. Reffin, A. Robertson and D. Weir. 2019. "Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts." *Studies in Conflict & Terrorism* 42, No. 1–2, pp. 141–60. https://doi.org/10.1080/1057610X.2018.1513984. Milton, D. October 2016. *Communication Breakdown: Unravelling the Islamic State's Media Efforts.* Combating Terrorism Center at West Point United States Military Academy.

11 Conway, M. 2 January 2020. "Routing the Extreme Right – Challenges for Social Media Platforms." *The RUSI Journal* 165, No. 1, pp. 108–13. https://doi.org/10.1080/03071847.2020.1727157.

12 Ganesh, B. and J. Bright. 2020. "Countering Extremists on Social Media."

13 Macdonald, S. and A. Staniforth. 2023. *Tackling Online Terrorist Content Together.*

The report starts by investigating the cultural fault lines affecting the relationship between American platforms and European public authorities, looking specifically at how contrasting conceptions of the State influenced how stakeholders defined and negotiated their involvement in countering terrorism and violent extremisms online. Although these tensions severely affected role perceptions, the second part shows that technological solutionism ultimately bridged the gap between public actors and the social media corporations studied, as can be seen in their shared enthusiasm for 'automated' moderation. The third part then analyses the strategies deployed by stakeholders to build working relationships without relinquishing their own preferences. These strategies amount to what can be regarded as a form of state-platform diplomacy.

RESEARCH DESIGN AND METHODOLOGICAL CONSIDERATIONS

In a qualitative, interpretative approach,[14] this report mobilises two main types of empirical materials that were collected in the course of a PhD thesis: a corpus of public documentation produced by the actors studied, and semi-structured interviews with some of the stakeholders involved.

First, a comprehensive corpus was collected of Facebook, Google and Twitter's corporate communications on terrorism up until late 2021, gathering all the public statements made by the companies or their representatives (including through the GIFCT) in which the terms "terrorism" and/or "violent extremism" were used. A collection of 273 documents, estimated at more than 2,600 pages of text, was compiled,

14 Devin, G. and Durand, M.-F. 2016. Chapitre 1 – Décrire, représenter, interpréter. In *Méthodes de recherche en relations internationales.* Paris: Presses de Sciences Po, pp. 15–38. https://doi.org/10.3917/scpo.devin.2016.01.0015.

curated and archived.[15] These span multiple formats and communication types, including blog posts, published interviews with corporate spokespeople, internal documents that were made public, content rules, transparency reports, parliamentary hearings and other documents, such as publicly available videos of events at which the companies were represented.

In parallel to the corpus collection and analysis,[16] a series of semi-directive interviews was conducted between late 2018 and 2023 with stakeholders from the public, private and civil society sectors (n=31), including current or former representatives from Facebook, Google, Twitter, GIFCT, the French Ministry for Europe and Foreign Affairs and Ministry of the Interior, the German Ministry of the Interior, the European Parliament, the European Commission, Tech against Terrorism, Moonshot CVE, and the Institute for Strategic Dialogue (ISD). Interviews lasted on average around one hour, and were conducted either face to face or online, depending on the logistical and time constraints on respondents. Conversations were recorded and transcribed when respondents had explicitly allowed this.[17] The interviews focused on the involvement of social media corporations in countering terrorism and violent extremisms, the public-private relations that this generates, and respondents' perceptions of these developments.

15 The precise size of the corpus is not known because of the many formats involved, including text, audio, videos, web pages, dynamic web pages, etc. This estimate was obtained by adding the page count of all the long, text-based sources in the corpus (those having 10 pages or more). It is therefore a conservative estimate, as videos without transcripts, podcasts and dynamic web pages are not included.

16 Preliminary findings from a discourse analysis of the corpus, and further information on how it was constituted, are available in: Borelli, M. 2024. "Countering 'terrorism' on social media: the use of a political category in the discourses of Meta, Google and Twitter." *Mots, les langages du politique* 134, pp. 57–79. https://doi.org/10.4000/mots.32949.

17 Note-taking was used in one instance, when permission had not been granted.

The corpus and interviews both fed into an inductive research design, where the iterative process of analysing public documents and stakeholder interviews helped to identify the meanings actors ascribed to their activities and practices, and enabled the production of 'thick descriptions' that could be used to reconstruct past events and identify drivers of change. On the one hand, interviews provided context for the information contained in the corpus, and allowed for the discussion of hypotheses derived from its analysis. On the other hand, the corpus helped determine which respondents were approached, and the questions they were asked. This framework is broadly inspired by the socio-historical approach developed in francophone International Relations scholarship.[18]

Within this context, this report specifically mobilises interview materials to investigate the lived experiences of early public-private cooperation surrounding TVEC in the EU, looking at how stakeholders across the public-private divide negotiated their involvement and navigated power relations to establish a satisfactory division of political labour. Interviews and quotes are cited in accordance with respondents' wishes,[19] which explains the varying levels of information available about each interview.

Because the study was conducted from France, it is the EU context, and in particular the French context, that informs this report. More than pure coincidence, however, this focus is also serendipitous. Because it was particularly affected by ISIS attacks and the recruitment of foreign fighters, France has been at the forefront of efforts to counter Daesh propaganda, quickly identifying mainstream social media corporations as strategic nodes on which

18 Devin, G. 2013. *Sociologie des relations internationales.* Paris: La Découverte; Devin, G. (ed.). 2016. *Méthodes de recherche en relations internationales.* Paris: Presses de Sciences Po.

19 Given the sensitivity of this research topic and the varying confidentiality requirements of the professionals interviewed, respondents were given a choice in specifying both how the interview would be credited (i.e. full attribution including name, role and institution; or partial attribution with just role and institution, or just institution; or full anonymisation), and the procedures for using verbatim quotes, if they could be used (i.e. with prior validation; or paraphrasing, etc.).

to exert pressure to curb the group's presence online. In response to the Paris terror attacks in 2015, these efforts were initially led by the Interior Ministry, through the Groupe de contact permanent, which comprised French authorities and the largest of the American platforms. This operational response was prolonged at the EU level, where the EU Internet Forum (EUIF) and Europol's Internet Referral Unit (EU IRU) were established.[20] The issue of hostile – and, in particular, terrorist – uses of the Internet then became a priority for French foreign policy during Emmanuel Macron's presidency, which saw the passing of the TCO Regulation and the Digital Services Act in the EU, as well as the establishment of the Christchurch Call with New Zealand. Because of its longstanding involvement alongside various other partners, most notably the United Kingdom, Germany and New Zealand, France therefore played a significant role in shaping the global governance regime on TVEC.

The global scale of the platforms, however, and the diffuse, transnational nature of terror-related threats, both contribute to a blurring of the various boundaries and scales involved, giving these initial negotiations a reach that exceeds the governance

20 Vieth, K. 2019. "Policing 'Online Radicalization': The Framing of Europol's Internet Referral Unit." In *Research Handbook on Human Rights and Digital Technology: Global Politics, Law and International Relations*, edited by B. Wagner, M.C. Kettemann and K. Vieth. Research Handbooks in Human Rights. Cheltenham: Edward Elgar Publishing, pp. 262–287; Bellanova, R. and M. de Goede. 2022. "Co-Producing Security: Platform Content Moderation and European Security Integration." *JCMS: Journal of Common Market Studies* 60, No. 5, pp. 1316–1334. https://doi.org/10.1111/jcms.13306.

levels analysed here. Indeed, EU-level legislation potentially entails global consequences in the form of a "Brussels effect",[21] while for the private sector, any platform-wide measure is necessarily a global one.[22]

21 Bradford, A. 2020. *The Brussels Effect: How the European Union Rules the World.* Oxford: Oxford University Press USA.

22 It should be noted, however, that while platform policies are global, their implementation is uneven and varies greatly depending on region and language. See in particular Gillespie, T. 2018. *Custodians of the Internet;* Common, M.F. 2020. "Fear the Reaper: How Content Moderation Rules Are Enforced on Social Media." *International Review of Law, Computers & Technology* 34, No. 2, pp. 126–52. https://doi.org/10.1080/13600869.2020.1733762. See also revelations by Facebook whistleblower Frances Haugen.


CULTURAL
FAULT LINES

IN ADDITION TO the public-private boundary, the relationship between European authorities and Facebook, Google and Twitter is marked by two other cultural differences. One has to do with the American, minimalist conception of the role of the state in society and the market, while the second concerns the particular ideology of Silicon Valley and its historically defiant view of (state) power. These differences influenced how European public actors and the US-based platforms defined their respective roles in the governance of terrorist content online – which, in turn, affected initial negotiations on the division of political labour on this issue. The following subsections explore those cultural fault lines, focusing on three of their repercussions: conflicting claims to legitimacy, with both public and private parties seeing themselves as best placed to protect their constituencies of citizens and users from the online repercussions of terrorism, and as being responsible for safeguarding these users' rights to freedom of expression and privacy; issues of reputation, as platform companies sought to distance themselves from the authorities in order to avoid accusations of censorship and collusion with state security apparatuses; and lastly, the influence of the US government, whose conception of free speech and economic interests served as a backdrop to negotiations.

CONFLICTING LEGITIMACY CLAIMS

Google, Facebook and Twitter are products of the particular Californian ecosystem of the Silicon Valley and its 'liberal-libertarian' political ideology, operating a synthesis between the libertarian counter-cultural ideals at the heart of the Internet's political project, and (neo)liberal capitalism.[23] These companies "are both financial groups seeking to maximise their profits and organisations that display an ethos promoting the emancipation of individuals

23 Loveluck, B. 2015. *Réseaux, libertés et contrôle: une généalogie politique d'internet*. Paris: Armand Colin.

and the defence of freedom of expression".[24] From the ideals of Internet pioneers evocative of John Perry Barlow's "Declaration of independence of cyberspace", they notably take on the conception of technology as a means of emancipating individuals from the constraints of traditional power structures.[25] And as actors who develop these tools and place them in the hands of civil society, they see and market themselves as counter-powers to states. Their transparency reporting, for instance, began as an effort to raise awareness of state overreach by publishing data on government requests. A Twitter blog post from 2019, for instance, states:

> *Transparency is not just the responsibility of tech companies. Governments and regulators should be transparent about their own actions, enabling people to know if content has been removed because of a decision Twitter made, or because of a government request. This transparency is essential if we are to foster an informed debate and mitigate the risk of inappropriate use of state power.*[26]

Similarly, a long post published by Mark Zuckerberg on his own Facebook profile in late 2018 reads: "I believe the world is better when (...) traditional gatekeepers like governments and media

24 Badouard, R. 2017. *Le Désenchantement de l'Internet: Désinformation, Rumeur et Propagande.* Présence. Questions de Société. Limoges: Fyp éditions, p. 93.

25 Haupt, J. 2021. "Facebook Futures: Mark Zuckerberg's Discursive Construction of a Better World." New Media & Society 23, No. 2, pp. 237–57. https://doi.org/10.1177/1461444820929315.

26 Gadde, V. 9 May 2019. "Key Data and Insights from Our 14th Twitter Transparency Report." Twitter Blog (blog). https://web.archive.org/web/20211022141105/https://blog.twitter.com/en_us/topics/company/2019/key-data-and-insights-from-our-14th-twitter-transparency-report.

companies don't control what ideas can be expressed".[27] This image of the Silicon Valley giants had its moment in the early 2010s during the Arab Spring, before gradually eroding in the latter half of the decade, beginning with the Snowden revelations and continuing with the series of scandals that constituted the so-called 'techlash'.[28] Nevertheless, Big Tech services remain hugely popular with European consumers, granting these 'superstar firms' a particular, enduring form of corporate power over the authorities seeking to regulate them.[29] Unsurprisingly, their marketing as guarantors of their users' freedoms in the face of state abuse is a source of frustration for European public actors, as expressed by a representative of the German Ministry of the Interior:

> *What is really quite extraordinary in this regard is that many large tech companies, that feed off of selling and exploiting user data, have more or less successfully portrayed themselves in society as guardians of the rights of their users, in opposition to 'bad states'. That has impacts on the public debate both on TCO and also on e-evidence and other regulations that affect the relationship between*

27 Zuckerberg, M. 28 December 2018. "A Blueprint for Content Governance and Enforcement." Facebook. www.facebook.com/notes/751449002072082.

28 Common, M.F. 2020. "Fear the reaper"; Badouard, R. 2017. *Le désenchantement de l'Internet*.

29 Woll, C. 2019. "Corporate Power Beyond Lobbying." *American Affairs* III, No. 3, pp. 38–55; Culpepper, P.D. and K. Thelen. 2020. "Are We All Amazon Primed? Consumers and the Politics of Platform Power." *Comparative Political Studies* 53, No. 2, pp. 288–318. https://doi.org/10.1177/0010414019852687.

> *states and tech companies. The social contract is the other way around though: it is the State who is charged with upholding the rights of its citizens.*[30]

Along the same lines, respondents from the French public sector expressed awareness of the reputational risks run by platforms if they associate themselves publicly with the authorities. One respondent from the French Ministry of Foreign Affairs (MFA) explained: "These are actors who have absolutely no interest in showing that they are collaborating with states, because there's [this whole spectrum] of Big Brother that is really very close to them. So they have objective interests in not doing so."* This is even more the case in the realm of countering violent extremisms (CVE) online, where many states suffer from a legitimacy deficit coming out of the post-9/11 era, after repeatedly invoking counter-terrorism to single out Muslim communities or to justify the indiscriminate surveillance of online communications.[31] In some respects, platforms have a greater capacity to be innovative in the policy area of CT/CVE, because they do not carry the same reputational "baggage", as argued by one respondent from a civil society organisation specialising in countering violent extremism:

> *Governments have a twofold problem: one, lack of money available to invest in this stuff, and then two, a lot of baggage around this, because they've done a lot of stuff previously and it's been ineffective,*

30 Where no source is given for a quoted passage, it has been taken from an interview conducted (and, where applicable, translated) for the purposes of this research. Interview quotes followed by an asterisk (*) have been translated from French.

31 Tréguer, F. 2017. "Pouvoir et résistance dans l'espace public: une contre-histoire d'Internet (XVe-XXIe siècle)." Paris: École des Hautes Études en Sciences Sociales. https://halshs.archives-ouvertes.fr/tel-01631122/document.

or it's made the problem worse. So, in terms of how liberated the private sector are, they are a lot more liberated to go into this space and do some more innovative work, but they have to balance that innovation with turning a profit.

REPUTATIONAL ISSUES

When it comes to the governance of online terrorist content, these reputational stakes influence corporate discourse in two ways. First, platform representatives insist on their ability to 'speak truth to power' by systematically evaluating the legality of takedown orders, and highlighting their own capacity to push back against abusive requests. Along similar lines, transparency efforts are frequently emphasised, in particular the data regularly published on how many injunctions they receive, some of which are made public on the Lumen database. A respondent from Google highlighted thus the risks posed by the TCO Regulation and its one-hour takedown delay:

*The new European regulation currently under discussion raises questions of freedom of expression, with the possibility of an administrative authority notifying us of content to be deleted, without any possibility of response for us, and without the time to evaluate the request because of the deadlines imposed. And when we say this to the Ministry of the Interior, they reply: "We're not asking you to look at the content, but to delete it", which is questionable from a legal standpoint, all the more so as everything goes through an administrative authority and not a judge... And even more so if this type of legislation is destined to be extended to hate speech.**

Private-sector respondents also stressed that, while the legality of government takedown requests rarely arises as an issue in France, on the one hand, the authorities are not immune to occasional errors of judgement, and on the other, it is imperative for companies to retain the capacity to push back against potentially abusive requests in the non-democratic states they operate in. From this perspective, the TCO Regulation sets a dangerous precedent in a global context, and its enforcement by European states with illiberal tendencies is worrisome. There is therefore a certain power struggle here between companies and European authorities, who, on the contrary, contend that the role of the private sector is limited to enforcing state requests, and that platforms have no legitimate entitlement to defend citizens from their own authorities. Recent interview-based research by Stuart Macdonald and Andrew Staniforth on the cooperation between tech companies and law enforcement points to this issue as an enduring one: when asked about current challenges and priorities for further progress, law-enforcement respondents emphasised the private sector's "delays in resolving requests",[32] while private-sector respondents raised concerns about "referrals for content that is only tenuously connected to terrorism or is not connected to it at all".[33]

Secondly, Silicon Valley culture and the reputational risks entailed in working with governments also translate to a strong preference for self-regulation, or at least, the *appearance* of self-regulation, even if this does not rule out a certain role for the state. A Google representative explains, for instance:

> *For me, the GIFCT and our collaboration with the rest of the industry on the hash database really illustrates that self-regulation is the best way to manage these issues effectively, because in any case, the sector is so much in flux that any regulation is bound to be obsolete within two years. On the*

32 Macdonald, S. and A. Staniforth. 2023. *Tackling Online Terrorist Content Together*, p. 14.

33 *Ibid.* p. 16.

> *other hand, it's true that the threat of government regulation is a driving force behind self-regulation, because it makes companies aware that there's a problem and that they need to take action. In contrast, state regulation often raises the question of abuse of power.**

This preference for self-regulation is also a source of incomprehension for French authorities, who do not necessarily find it justified. Indeed, it can be considered less risky politically for firms to limit their CT/CVE efforts to what is legally required of them in the various jurisdictions they operate in. The former French ambassador for digital affairs, David Martinon, recalls his talks with Google, Twitter and Facebook about terrorism in the following manner:

> *They were never opposed [to what we asked them to do], but they felt it was up to them to do it, and not up to us to tell them what to do. It was even a rather obstinate position, as it was out of the question for them to admit their weaknesses or mistakes. It was up to them to make their own policy. And so it was very, very often a 'dialogue of the deaf'.**

M. Martinon tellingly uses the French idiom "a dialogue of the deaf" to characterise their negotiations, emphasising the difficulty encountered by his team in establishing a meaningful dialogue with the companies. Similar communication problems were reported by other public-sector respondents, while parliamentary hearings often made them publicly noticeable, as exemplified by the following exchange between a Google representative and a UK Member of Parliament (MP) during a Home Affairs Select Committee hearing on YouTube's moderation of National Action content. In this excerpt, the

MP had to ask a factual question eight times before finally obtaining a precise response:

> *Member of Parliament (MP): You opened your statement saying that there are three actions that you've taken to improve this [moderation of National Action]. When did they start?*
>
> *Google representative (GR): Immediately.*
>
> *MP: From when?*
>
> *GR: From when the videos were escalated to us, we could clearly see there was a volume…*
>
> *MP: So tell me, when did these start to go to specialised reviewers?*
>
> *GR: Immediately.*
>
> *MP: On what date?*
>
> *GR: Now.*
>
> *MP: No, immediately is when you asked for it to be done.*
>
> *GR: Right, and it's been done.*
>
> *MP: No, but when is immediately? When did you ask for it to be done? Was it today? Was it Friday? Was it a month ago? Or six months ago?*
>
> *GR: When the Chair brought these videos to our attention…*

MP: They've been brought to your attention for the last 18 months, so when in that 18 months did they start to go to specialised reviewers? When did you get improved training for your reviewers? And when did you fine-tune your technology?

GR: We are putting those…

MP: No, no. You say they are in place, and they were in place immediately. So I'd like to know the date of those three improvements, because you've told this Committee, you're on the record as saying they had been done, and I'd like to know when they started.

GR: These videos, National Action…

MP: No, these three points, these three improvements that you have put in place, when did they start?

GR: Late last week.

MP: Wow.[34]

34 UK Home Affairs Committee. "Policing for the future" evidence session – Tuesday 13 March 2018 Meeting started at 3.05pm, ended 5.57pm – Witnesses: William McCants, Global Policy Lead for Counterterrorism, YouTube, Google. Parliamentlive.tv, 2018. www.parliamentlive.tv/Event/Index/2bc566ce-1ae1–48c4-ac20–00555c46a4b9.

THE LONG SHADOW OF THE US GOVERNMENT

For all their marketing on 'speaking truth to power', the historic proximity of Silicon Valley firms to United States (US) authorities and interests abroad has been well documented.[35] On the issue of TVEC too, Facebook, Google and Twitter's preference for self-regulation aligns with the US government's stance on the matter. Indeed, in the US, a good deal of what is considered illegal terrorist content in the EU is protected under the First Amendment of the Constitution.[36] To manage this so-called "lawful but awful" content[37] on social media, US authorities therefore rely on the policies voluntarily developed and implemented by platform companies. Meanwhile, on the world

35 Political economy scholarship holds that, in general, multinational corporations tend to act as ambassadors for their home countries, prolonging their power on the world stage. See, in particular, Strange, S. 1996. *The Retreat of the State: The Diffusion of Power in the World Economy*. Cambridge: Cambridge University Press. Social media corporations especially maintained very close ties with the Obama administration. See for instance Assange, J. 2014. *When Google Met WikiLeaks*. New York & London: OR Books. https://doi.org/10.2307/j.ctt1bkm5qf; and Thibout, C. 2021. "Google et l'État fédéral états-unien: interdépendance, contestation et hybridation." *Entreprises et histoire* 104, No. 3, pp. 142–63. https://doi.org/10.3917/eh.104.0142.

36 Neumann, P.R. 2013. "Options and Strategies for Countering Online Radicalization in the United States." Studies in Conflict & Terrorism 36, No. 6, pp. 431–59. https://doi.org/10.1080/1057610X.2013.784568.

37 Keller, D. 28 June 2022. "Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users." *The University of Chicago Law Review* Online Archive. https://lawreviewblog.uchicago.edu/2022/06/28/keller-control-over-speech.

stage, the US's formal position[38] on the issue of TVEC is characterised by a general preference for self-regulation by private actors, as well as for 'softer', counter-narratives-based approaches[39] rather than content takedowns. This was expressed at the 2019 Internet Governance Forum by a representative from the State Department's Bureau of Counter-Terrorism, whose words bear a striking resemblance to those of corporate representatives:

> *Our experience is that, and we continue to contend, that voluntary collaboration with the technology companies and all stakeholders on this issue is a better approach. We think that the companies know better the content on their platforms, how to identify and remove content more quickly and to keep it from propagating. And we continue to argue that some of the regulations we're seeing (...) can in fact be an inspiration for more repressive regimes to fine or imprison company executives for example for not removing 'extremist' content that may actually be political dissent.*[40]

38 While this is the formal US position, however, the Twitter Files have confirmed the existence of an informal system of cooperation between private social media companies and US law enforcement with regard to 'legal but harmful' content. Emails made public after Musk's takeover of Twitter show that FBI officers regularly brought violative content to the attention of platform employees, in a system that seemed to function as an 'informal IRU', renewing debates in the US about "jawboning" (see Lakier, G. 26 July 2021. "Informal Government Coercion and The Problem of 'Jawboning.'" *Lawfare* (blog). www.lawfareblog.com/informal-government-coercion-and-problem-jawboning.) While the particular exchanges revealed in the Twitter Files concern disinformation, it is likely that such a system also extends to violent extremist and terrorist content.

39 See B. Ganesh and J. Bright. 2020. "Countering Violent Extremists on Social Media."

40 IGF 2019 – Day 2 – Convention Hall II – Addressing Violent Extremist Content Online – Floor. Berlin, 2019 www.youtube.com/watch?v=PjjMXoYFxaE.

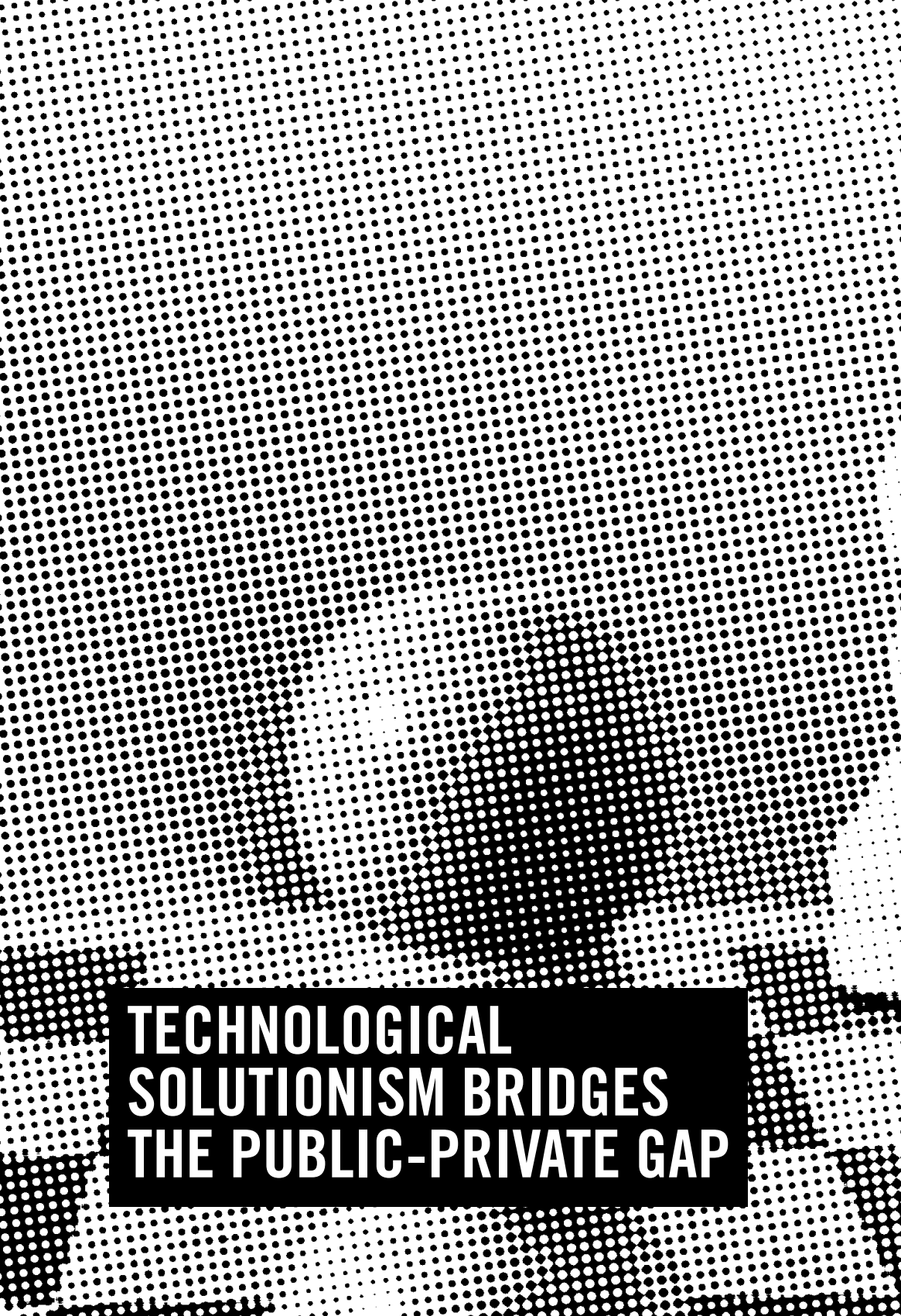
The satisfaction of US officials with the tech industry's self-regulation on terror-related threats appears somewhat hypocritical, given the calls for state guidance from platform representatives themselves,[41] and the key role the EU has played in bringing American corporations to take action on this issue. At any rate, the plea from France, Germany and the UK for regulation within the European Union was viewed on the other side of the Atlantic with suspicion and scepticism. While negotiations on what would become the TCO Regulation progressed in the EU, US authorities had a more favourable view of the parallel United Nations Counter-Terrorism Executive Directorate (UN CTED)-led process on TVEC, owing to its strictly voluntary approach.[42] Both in bilateral relations between France and the United States, and in multilateral forums such as the G7, where France has been a vocal advocate on this issue, the shadow of the US government has loomed large over the establishment of a governance regime for terrorist content. Yet, although the US was reluctant to accept the prospect of 'hard' regulation, respondents from the French MFA did note a gradual evolution in its position, against a background of Daesh claiming attacks on American soil and the growing threat of violent right-wing extremism. Evidence of this growing preoccupation can be seen in the fact that Facebook, Google/YouTube and Twitter were summoned by three different congressional committees between late 2017 and 2019 to testify publicly about their work on

41 Caplan, R. 2018. *Content or Context Moderation?* Data & Society. https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf.

42 The UN CTED convened voluntary talks with industry on the issue of terrorist content online as early as 2014, in the context of Security Council discussions on foreign terrorist fighters. These have notably led to the establishment of the Tech against Terrorism public-private partnership.

terrorism.[43] However, it was not until the aftermath of the Capitol Riots of January 2021 that the United States joined the Christchurch Call, thereby finally becoming an official part of the emerging global regime to counter terrorism and violent extremisms online.

43 US Senate Committee on the Judiciary. 2017. Extremist Content and Russian Disinformation Online: Working with Tech to Find Solutions. www.judiciary.senate.gov/meetings/extremist-content-and-russian-disinformation-online-working-with-tech-to-find-solutions.; US Senate Committee on Commerce, Science, & Transportation. 2018. Terrorism and Social Media: #IsBigTechDoingEnough? www.commerce.senate.gov/2018/1/terrorism-and-social-media-isbigtechdoingenough.; US House Committee on Homeland Security. 2019. Examining Social Media Companies' Efforts to Counter Online Terror Content and Misinformation. https://web.archive.org/web/20221209123356/https://homeland.house.gov/activities/hearings/examining-social-media-companies-efforts-to-counter-online-terror-content-and-misinformation.


TECHNOLOGICAL
SOLUTIONISM BRIDGES
THE PUBLIC-PRIVATE GAP

DESPITE THESE CULTURAL fault lines, European public actors and companies alike sought efficiency, first and foremost, in their efforts to take down ISIS propaganda online or prevent it from appearing in the first place. Their conceptions of efficiency differed slightly, with public actors prioritising the *speed* of takedowns while private actors pointed to the *reach* and *virality* of content. Both sides, however, shared a common interest in developing and deploying technologies to manage terrorist content on social media.[44] Despite their many limitations,[45] automated moderation technologies were seen by respondents as an asset that gave platforms undeniable value in the field of counter-terrorism by enabling effective action on a global scale, and in record time. Conceptions continued to diverge, however, on the role of platform companies in the promotion of counter-narratives on their services.

THE DEVELOPMENT AND DEPLOYMENT OF AUTOMATED MODERATION TOOLS

Since their inception, companies like Google, Facebook and Twitter have been known for their techno-solutionism.[46] However, public entities also seem to have adopted some of these narratives, in particular when it comes to TVEC. The UK Home Office, for

44 Macdonald, S., S. Giro Correia and A.-L. Watkin. 2019. "Regulating Terrorist Content on Social Media: Automation and the Rule of Law." *International Journal of Law in Context* 15, No. 2, pp. 183–97. https://doi.org/10.1017/S1744552319000119.

45 See in particular Gillespie, T. 2020. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7, No. 2, pp. 1–5. https://doi.org/10.1177/2053951720943234. Gorwa, R., R. Binns and C. Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7, No. 1, pp. 1–15. https://doi.org/10.1177/2053951719897945. Ganesh, B. 16 March 2021. "Platform Racism: How Minimizing Racism Privileges Far Right Extremism." *SSRC Items* (blog). https://items.ssrc.org/extremism-online/platform-racism-how-minimizing-racism-privileges-far-right-extremism.

46 Morozov, E. 2013. To Save Everything, Click Here: *The Folly of Technological Solutionism*. First edition. New York: PublicAffairs.

instance, even went so far as to develop its own AI tool for identifying ISIS propaganda, and made it available to all platforms.[47] Initially, public actors were even 'more royalist than the king', displaying more faith in these technologies than the tech companies themselves, and they used it to push the firms to take action. Authorities repeatedly referred to the PhotoDNA system for countering Child Sexual Abuse Material (CSAM) when pushing for similar efforts in the field of counter-terrorism, in a case of "censorship creep".[48] In May 2016 for instance, in the context of a parliamentary enquiry on Daesh's capabilities, a French MP put the following question to representatives from Twitter, Facebook and Google:

> *You spoke of an algorithm that identifies child pornography content. Why is it that no such algorithm exists for terrorism? Intellectually, it doesn't seem any more complicated to me.** [49]

At the time, none of the firms studied had deployed TVEC-specific systems on their services, and the corporate representatives found themselves in the unusual position of explaining to MPs why, in fact, such technology would not work for terrorism, because of the contextual understanding necessary and the human rights risks involved. Despite this, the three firms would announce the launch of the Shared Industry Hash Database (SIHD) with Microsoft later

47 Home Office, and A. Rudd. 13 February 2018. "New Technology Revealed to Help Fight Terrorist Content Online." GOV.UK. www.gov.uk/government/news/new-technology-revealed-to-help-fight-terrorist-content-online.

48 Citron, D.K. 2018. "Extremist Speech, Compelled Conformity, and Censorship Creep."

49 Arif, Kader. 13 July 2016. *Rapport d'information fait au nom de la Mission d'information sur les moyens de Daech, Tome II*. Paris: Assemblée Nationale, p. 497.

the same year, in December 2016.[50] Since then, the firms have remained somewhat cautious in their public communications about the technology, careful not to make it look 'too easy', perhaps for fear of what authorities would ask of them next.[51] When companies did introduce automated systems for detecting terrorist content on their respective platforms, they were careful to keep managing expectations.[52] Twitter representatives, for instance, repeated multiple times that "there is no 'magic algorithm' for identifying terrorist content".[53] The CSAM comparison[54] kept being used by

50 The SIHD is a database of terrorist-content hashes shared among various GIFCT members. When a company identifies a piece of content on its services that qualifies for inclusion in the database, it can hash it and upload the hash onto a shared platform. Other members of the consortium can then use the hashes to scan their own services for the content, taking it down if it breaches their terms of service, and/or adding it to their upload filters. The SIHD was modelled on the system for countering child sexual abuse materials (CSAM) based on PhotoDNA technology developed by Microsoft. The rationale behind it is that content that has been taken off one service should not be able to reappear on another. Up until the Christchurch attack, the operation of the SIHD was based on UN designation lists.

51 Although public-sector respondents prioritised terrorist content in their talks with the platform companies at the time, they often saw a continuum between this type of content and hate speech and disinformation.

52 Crosset, V. and B. Dupont. 2022. "Cognitive Assemblages: The Entangled Nature of Algorithmic Content Moderation." *Big Data & Society* 9, No. 2, pp. 1–13. https://doi.org/10.1177/20539517221143361.

53 See Monje, C. 17 January 2018. "U.S. Senate Committee on Commerce, Science, and Transportation – Hearing on Terrorism and Social Media – January 17, 2018 – Statement of Carlos Monje, Jr., Director, Public Policy & Philanthropy, U.S. & Canada, Twitter, Inc." US Senate Committee on Commerce, Science, and Transportation. www.commerce.senate.gov/services/files/447E3321–1215–47BC-93EA-AD295CF16D80; and Edgett, S. 19 January 2018. "Edgett Responses to Questions for the Record." US Senate Committee on the Judiciary. www.judiciary.senate.gov/imo/media/doc/Edgett%20Responses.pdf.

54 The CSAM comparison also acts as a rhetorical device to depoliticise the notion of 'terrorist' content. It is used by all actors to justify their focus on this category of 'undesirable' content, categorising it as both a more objectively urgent harm to tackle, and an 'easier' case than that of hate speech and disinformation. See Borelli, M. 2024. "Countering "terrorism" on social media."

public actors to push the firms into action, including during the TERREG negotiations at the EU Parliament a few years later, as recounted by Daniel Dalton, a UK MEP who was the Rapporteur for the text:

> *Now, they will probably have to apply a similar approach to terrorist content, that they applied to CSAM. This was a point I always made to them: if you can block most of that material – we're not getting huge complaints about that type of material, so therefore you're obviously able to identify it, block it and stop it coming back – why can't you do that on terrorist content? And that was really the sort of point that they'd take.*

Once automated tools for moderating terrorist content were deployed, European authorities seemed relatively satisfied with them, while nonetheless noting variations in their efficiency both by language and on the different services owned by the firms. Facebook, for example, was known to be better moderated than Instagram, and YouTube than Google+ and Drive, while English-language content was known to be better moderated than that in French or German. A respondent from the French MFA expressed satisfaction with the automation put in place by the three firms, although they hinted that it could come at a price for free speech, particularly on Twitter, which could invest only limited resources:

> *Facebook and Google, and in this case YouTube, were showing that they were making efforts, but also because they have unlimited resources, (...) the automation part was pretty convincing, but not quantifiable. (...) I was confident that they were indeed making efforts in terms of investment*

> *in artificial intelligence so that terror content could be taken down quickly (…) The problem with Twitter was that they couldn't do what the others were doing because they didn't have the same resources (…) they did not have the means, so they were sometimes indiscriminate in their deletions.**

Despite their confidence overall, public-sector respondents also expressed a certain impatience with the firms over the opacity of their automated systems and the metrics used to measure their efficiency. Frustration over the unverifiable nature of figures put forward by platform executives was a recurring theme among public service respondents. Former French Ambassador for Digital Affairs David Martinon recalled the results of terrorist content moderation presented by platform representatives:

> *They would bombard us with figures, each one more miraculous than the last. Except that these were* ***their*** *figures, and they gave us no opportunity to corroborate them. For example, 99% of the content uploaded to YouTube or Facebook is screened in less than 10 seconds. (…) It's always 99%, and always has been 99%.**

Likewise, the GIFCT's hash database was praised as a promising tool by public-sector respondents, but its opacity was criticised:

> *This hash database, in principle, is a very good idea, but we don't really know how it works in practice. We have never managed to get a perfectly clear answer from the major platforms regarding the functioning of this database. We don't know if, as*

> *soon as a platform removes content, it automatically adds it to the database. Does removed content automatically get shared with the other members of the consortium? We don't know if they really use the database to scan their platforms proactively based on what has already been identified. We are given figures on the number of digital fingerprints in the database, but we actually don't know anything.**

Despite this friction, automated moderation systems, as forms of data-driven policymaking, are clearly interesting to public authorities, who see image-, video- and text-recognition technologies as a means of achieving large-scale efficiency – perhaps even of seeing their own speech norms applied extraterritorially – while at the same time reducing takedown delays. At the EU level, this satisfaction with the systems put in place by the major platforms was also reflected in the first version of the TCO Regulation proposed by the Commission, which France was heavily involved in drafting and which contained an obligation for platforms affected by TVEC to filter terrorist content actively. The French Ministry of the Interior saw the mandatory automation of terrorist content moderation as one of the most important provisions within the text:

> *What is at stake in the TERREG is the 'golden hour', i.e. withdrawal [of notified content] within one hour. It's also being able to penalise [companies] when removal orders aren't systematically respected, (…) and then, it's also the implementation of proactive measures, i.e. obliging them to set up automatic tools for detecting terrorist content in order to prevent, for example, the reappearance of content that has already been reported (…). In fact, the problem is that reported content is removed, but*

> *people have time to download it, so it's disseminated from platform to platform. The idea is to remove content permanently.**

Behind automated moderation technologies, governments also see the possibility of compliance with the so-called 'golden hour', i.e. the deadline by which platforms are supposed to remove terrorist content whose takedown has been ordered by national referral authorities. This one-hour deadline is also one of the key provisions of the legislation. Authorities argued that the longer content remains online, the more likely it is to be downloaded, modified and disseminated in various forms. Both the 'golden hour' and the mandatory use of "proactive measures" proposed in the original draft of the Regulation have been heavily criticised by digital and human rights organisations, who succeeded in getting the latter provision dropped in the final version of the text.[55] Despite concerns that the one-hour time limit would lead to over-blocking, and would impose an undue burden on smaller companies, it was retained, although its basis seems somewhat arbitrary. The Christchurch attack, for example, was broadcast on Facebook Live for 'only' 17 minutes, yet this quarter of an hour was enough time for the content to be downloaded, copied, modified and disseminated across the entire web. Within 24 hours of the attack, Facebook said it had blocked 1.2 million attempts to repost the video and deleted 300,000 modified versions of the livestream which had made it onto the platform.[56] The content also quickly spread to other services, becoming viral. Around ten days

55 Ahmed, R. 2023. "Negotiating Fundamental Rights: Civil Society and the EU Regulation on Addressing the Dissemination of Terrorist Content Online." *Studies in Conflict & Terrorism*, pp. 1–25. https://doi.org/10.1080/1057610X.2023.2222890.

56 Rosen, G. 20 March 2019. "A Further Update on New Zealand Terrorist Attack." Facebook Newsroom (blog). https://web.archive.org/web/20191125123121/https://about.fb.com//news/2019/03/technical-update-on-new-zealand.

after the attack, Vijaya Gadde, then director of Twitter's Legal, Policy and Trust & Safety teams, counted more than 300 variations of the video on Twitter, recalling:

> *It was basically a game of whack-a-mole: we would identify one particular version of this video, and we could use that technology to prevent further uploads of that, then the video would modify and we'd have to go through that process again.*[57]

In total, 800 distinct versions of the attack video were identified by GIFCT member companies, who for the first time added content unrelated to Daesh or al-Qaeda to their common database.[58]

Platform companies also saw their technical capacities as their main added value in the field of counter-terrorism. Corporate representatives emphasised governments' lack of capacity in this area, and their own proactiveness in these innovative developments. In their view, their deployment of automated moderation demonstrates their positioning as responsible companies, but it must remain their prerogative. The ability of the public sector to impose the use of these tools is viewed with scepticism, as evidenced in this quote from a Google respondent:

> *Our added value is what we do well, which is the technology space, for one. (…) we've got the engineers who know how to do it, we've got the content we've removed to be able to train the machines, and then we've got the platforms that can use the technology to identify new pieces of content that have been*

57 "Twitter's Vijaya Gadde Says Removing Videos of the New Zealand Shooting Was 'Difficult'". 2019. www.youtube.com/watch?v=mcNcborEbmE.

58 GIFCT. 18 March 2019. "Industry Cooperation." GIFCT (blog). https://gifct.org/2019/03/18/industry-cooperation.

> *posted. That's just all so squarely in the wheelhouse of a technology company (...) And in fact, you see things like the proposed EU Regulation about governments being able to mandate certain technologies, and you think actually that's something tech companies should do on their own, and can do better, know what is the right technology to go after this problem.*

THE PROMOTION OF COUNTER-NARRATIVES

But private-sector respondents also like to emphasise that their capacity for innovation and their contribution to the fields of CT and CVE extend beyond content moderation. Facebook, Google and Twitter saw themselves as having a responsibility to promote counter-narratives, by training civil society organisations active in this space, donating advertising credits and direct funding and/or amplifying the visibility of counter- or alternative narratives. According to a respondent from Facebook:

> *When we can help empower people, we have to do it. It's our responsibility to help them arm themselves digitally to face up [to violent extremism]. (...) And we have to help them do it because they know better than we do (...) what to say, which narratives work. So we lend them the Facebook megaphone to do it, but we're not going to define the messages ourselves. When it comes to counter-speech, a number of studies have shown that the messenger counts as much as the message: government counter-speech doesn't work, and our counter-speech wouldn't work*

> *any better, because we'd be seen as big American corporations in the pay of governments. We need to find the right megaphones.**[59]

Google and YouTube have likewise invested in this space by developing the ReDirect Method through their internal think tank, Jigsaw. This project, which originated from a visit to Iraq and Syria by Jigsaw teams to interview jihadists about the role of the Internet in their own radicalisation trajectories, involves redirecting users whose queries express an interest in terrorism to curated YouTube playlists of counter- or alternative narratives. ReDirect is now used by other actors, including Meta, and is deployed on topics other than terrorism, such as self-harm.

European public-sector respondents, on the whole, were less enthusiastic about the counter-speech work of platforms. One respondent from the German Interior Ministry even argued that the involvement of private firms in counter-narratives was problematic: "I would be careful about private initiatives about extreme political views. These are part of the political dialogue as long as they do not represent any criminal actions." More generally, these initiatives were generally dismissed as insignificant PR stunts by the French MFA, by partner services in charge of countering radicalisation, and by foreign colleagues. In particular they criticised the small scale of the initiatives, relative to how much firms communicated about them, as well as their strong focus on English-language content. Drawing a parallel with negotiations on disinformation, a respondent from the French MFA recalled, half-jokingly:

59 Facebook had conducted research on counter-narratives with the UK-based think tank Demos as early as 2015/2016. Its findings are notably reported here: Facebook's Top Content Judge Discusses Online Terrorism, 2016. www.youtube.com/watch?v=0Uxaf8JXzZ0.

> *This is really how they respond: "we've understood the problem", first element of response; "but it's a major issue", second element, "so we shouldn't rush into anything". OK, so what have you done? "We've developed this local initiative." And then they tell you about some initiative at a high school in Texas or somewhere (...). Fourth element: "so we're thinking about it, we'll see if we implement it". Of course, you come back three months later, and there's nothing. But it's not a big deal, because on the spot, they've shown you a whole bunch of measures they've taken. These are ultimately completely superficial, really completely, but it obfuscates the issue. So, they buy themselves time. In my opinion, their main strategy is to buy time to implement changes and announce them, and capitalise on them.**

Similar conclusions about the scale of initiatives were drawn by a respondent from the Institute for Strategic Dialogue (ISD), a civil society organisation which was involved in such programmes funded by the companies at the time:

> *I don't think it's ever been done at a scale where you'd genuinely be able to say you're proportionate to the problem. (...) Obviously there's a whole other set of problems about data access and transparency from companies, given that we still don't really know the true scale of a lot of these problems, so it's hard to assess what a proportionate response would look like (...). A lot of the work we did, we would often describe it as "testing out different approaches".*

Meanwhile, another respondent from the French MFA hinted that, in terms of countering radicalisation, authorities expected better design choices from platforms, rather than counter-narratives:

> *Algorithmic echo chambers are complicated because there are two sides to the question. There's the civil society side, promoting an alternative discourse, but there's also the very technical side, the functioning of algorithms. And on that, platforms have a role to play. I'm entirely open to the idea that platforms have no role in supporting civil society and that indeed, that is more within the government's remit. But on the question that escapes us, the potentially harmful functioning of algorithms, we will one day need to have a more open discussion with the platforms, because these are real issues.**

In light of similar observations over time, ISD also shifted its position towards supporting design-oriented regulation of the platforms as the most effective way to counter radicalisation at scale:

> *From 2013 to 2015–16, there was this sort of competition paradigm. (…) During that period, the main idea was: how do we compete? How do we offer an alternative? So, a more free speech conception. But increasingly, the more of that work that we tried to do, I think we increasingly realised that those online environments aren't necessarily a level playing field because of the way they're designed. Because of some of the features of those platforms, I think controversial or inflammatory content often has an advantage.*

Despite their lack of success with EU public officials, platforms' investment in counter-narratives does appear to align with the official stance of the US authorities. Indeed, in a 'marketplace of ideas'-inspired approach, the US maintains that the cornerstone of the fight against terrorism online is counterspeech, in contrast to a more repressive European approach, focused on takedowns.[60] The State Department representative at the 2019 IGF argued:

> *We continue to maintain that the most effective means to counter terrorist and other objectionable content online is not through censorship or repression but through more speech that promotes tolerance. We emphasize the importance therefore of credible alternative narratives as the primary means by which we can undermine and defeat terrorist messaging. (...) We also recognize that banning of speech can be counterproductive to our efforts. It can raise its profile. It can also drive it into darker places and in fact undermine our counter-terrorism effort.*[61]

These words echoed those of corporate representatives, for instance when Twitter's Senior Public Policy Strategist argued: "We will not solve problems by removing content alone. We should not underestimate the power of open conversation to change minds,

60 That is not to say that European public stakeholders denied the importance of counter-speech in the fight against violent extremism, or the rightful involvement of civil society in this area, alongside the CIPDR in France. However, counter-narrative work was viewed through a slightly different lens – as a part of efforts to combat racism or online hate, rather than as a counter-terrorism practice.

61 IGF, "Adressing Violent Extremist Content", 2019.

perspectives, and behaviors."[62] In fact, the proximity of corporate counter-narrative efforts to those of the US state apparatus is more than ideological. One of the first counter-narrative programmes in which Facebook was involved, the Peer to Peer: Facebook Global Digital Challenge, was initially developed by the US State Department, the Department of Homeland Security and EdVenture Partners before it was taken over by the social media firm in October 2016.[63]

62 Pickles, Nick. "Written Testimony Of Nick Pickles, Senior Public Policy Strategist, Twitter, Inc., Before The United States House Of Representatives Committee On Homeland Security Hearing Titled 'Examining Social Media Companies' Efforts To Counter Online Terror Content And Misinformation.' June 26, 2019. Unclassified." National Security Archive, 26 June 2019. https://nsarchive.gwu.edu/document/20059-national-security-archive-093-written-testimony.

63 "Recruiting College Students to Fight Extremists Online." PBS NewsHour, 30 January 2016. www.pbs.org/newshour/show/recruiting-college-students-to-fight-extremists-online.


STATE-PLATFORM
DIPLOMACY

TO DEFEND THEIR interests and establish effective cooperation despite obvious tensions and the power struggles underpinning their relations, European public authorities and American-based platforms have developed a form of "diplomacy", i.e. strategies for managing their alterity.[64] This third part of the report investigates some of those practices implemented "to prevent disagreements from turning into conflicts or, further upstream, to reduce the possibility of such disagreements"*[65] on the particular issue of terrorist uses of the Internet.

FRAMING THE ISSUE OF TERRORIST CONTENT ONLINE

First, the initial framing[66] of the issue of terrorist content online, as promoted by the French government and its European allies, was crafted strategically. Public-sector respondents overtly referred to a sovereignty frame: in the national public sphere, national legislation should be applied online as it is offline. The heart of the matter, then, was the deletion of content hosted by platforms that is illegal by virtue of national law, particularly so when it was referred by national authorities. This frame is strategic in that it has the effect of setting aside the complex and controversial topic of the actual role of social media platforms, their ranking and recommendation algorithms, in

64 Balzacq, T., F. Charillon and F. Ramel. 2018. *Manuel de diplomatie. Relations internationales*. Paris: Presses de Sciences Po.

65 *Ibid*., p. 16.

66 According to Robert Entman, "Framing entails selecting and highlighting some facets of events or issues, and making connections among them so as to promote a particular interpretation, evaluation, and/or solution." See Entman, R.M. 2003. "Cascading Activation: Contesting the White House's Frame After 9/11." *Political Communication* 20, No. 4, pp. 415–432. https://doi.org/10.1080/10584600390244176, p. 417.

processes of radicalisation into violence.[67] According to a respondent from the French MFA, corporate representatives were utterly unwilling to engage on the matter of echo chambers at the time:

> *I can't say we've arrived at very concrete answers regarding echo chambers because it's really a new subject. Furthermore, it touches on the core of what platforms are about – their algorithms. So, of course, they are not prepared to tell us how they work, nor are they willing to concede any alteration to the functioning of their algorithms, which for them would mean questioning their business model.**

A respondent from ISD concurred, noting that design changes were generally a red line for the firms:

> *Over the years my view of the companies has been that they want to be seen to be addressing these problems, but only to the extent that it doesn't impact their business model or the platform more broadly. There's a certain distance they will go, but there's also a place where they don't necessarily want to go, and I think that's where regulations have to come in to try and change those incentives.*

67 Conway, M. 2017. "Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research." *Studies in Conflict & Terrorism* 40, No. 1, pp. 77–98; Klausen, J., R. Libretti, B.W.K. Hung and A.P. Jayasumana. 2018. "Radicalization Trajectories: An Evidence-Based Computational Approach to Dynamic Risk Assessment of 'Homegrown' Jihadists." *Studies in Conflict & Terrorism* 43, No. 7, pp. 588–615. https://doi.org/10.1080/1057610X.2018.1492819.

The initial sovereignty frame adopted by European authorities in the early negotiations that led to the TERREG thus circumscribed discussions on terrorism online, restricting them to the application of national laws on these American services, via the identification and removal of illegal content. A respondent from the French MFA explains:

> *What we're criticising them for isn't so much that their platforms play a role in radicalisation, (...) what we're criticising them for is allowing speech that is in itself illegal to spread in the non-digital space. (...) in any case, the EU regulation we're working on today has much less to do with radicalisation than with the propagation of content.**

This was also emphasised by a respondent from the German Ministry of the Interior, who stressed that the role of platform companies is limited to the application of the rules laid down by the public authorities:

> *Laws that are applied in the public sphere of the analogue world must be applied here [online] too. It is the state's responsibility to legislate and also to enforce relevant laws. However, the execution is in many cases only possible in cooperation with private entities. We see this in other areas as well, such as money laundering. (...) Carrying out measures on the technical level must be done by the relevant companies playing by the rules the state works out. In other words, the main role of companies is one of a technical character.*

Lastly, the sovereignty frame was also seen by some respondents as a way of facilitating a general consensus at the European level, and thereby of obtaining a symbolic victory in the more general power struggle pitting European states against so-called Big Tech.[68] An analyst from the French MFA notes:

> *The state's ability to use this issue [terrorism] to achieve its own objectives should not be underestimated. As for the question of platforms, when we succeed in forcing Facebook to collaborate with us, it's good not only for the issue area at hand, it's also good to get Facebook to yield (...). The objective of obtaining something from them is almost as important as trying to reverse this power relation.**

HARNESSING THE TECHLASH AS A WINDOW OF OPPORTUNITY

The development of state-platform diplomacy on TVEC was also facilitated by the changing context affecting Facebook and Twitter, and to a lesser extent Google.[69] The Cambridge Analytica scandal and revelations of Russian foreign interference created a crisis of confidence in social media companies and a global context of 'techlash', to which Daesh's exploitation of their services only added. In this regard, public accusations levelled against platforms

68 Bellanova, R. and M. de Goede. 2022. "Co-Producing Security".

69 Many respondents agreed that, for various reasons, less scrutiny was applied to Google and its YouTube platform than what Facebook and Twitter experienced at the hands of the media and regulators. See also Douek, E. 17 November 2020. "Why Isn't Susan Wojcicki Getting Grilled By Congress?" *Wired*. www.wired.com/story/why-isnt-susan-wojcicki-getting-grilled-by-congress.

by political leaders after various attacks are particularly relevant.[70] An example was the advertisers' boycott of YouTube, following a journalistic investigation revealing that some major brands were indirectly funding Daesh and Britain First via ads on its channels.[71] It became imperative for firms to position themselves as responsible actors in order to regain the trust of advertisers and users, and also to pre-empt potentially costly regulation. Indeed, the acceleration in Germany (NetzDG), the UK (Online Harms) and France (Avia law) of legislative projects at the national level on hate speech and illegal content also added to the pressure on companies, and threatened the integrity of the EU's single digital market.[72] To start repairing their image, platform companies prioritised counter-terrorism – a field which, in the context of the ISIS threat, they anticipated would be consensual enough not to generate accusations of censorship.[73] As noted by an analyst from the French MFA:

70 See Borelli, M. 2021. "Social media corporations as actors of CT", on securitisation processes tying social media to terrorism in the French and British public spheres.

71 Mostrous, A. 9 January 2017. "Big Brands Fund Terror through Online Adverts." *The Times*, sec. news. www.thetimes.co.uk/article/big-brands-fund-terror-knnxfgb98.

72 On those laws and their various outcomes see Badouard, R. 2020. *Les Nouvelles Lois du web: Modération et censure*. Paris: Seuil. It is also noteworthy that these national laws also put pressure on the EU to act on content regulation or risk seeing the digital single market fragmented into a patchwork of different national (and potentially conflicting) legal frameworks. In this respect, EU institutions and platform companies shared an interest in avoiding fragmentation.

73 Initial 'counter-terrorism' efforts by the companies were largely targeted against Daesh. Research on Twitter has found in particular that the company's disruption of ISIS far outweighed that of other jihadist groups and far-right extremists. See, on ISIS vs. other jihadist groups: Conway, M., et al. 2019. "Disrupting Daesh"; and on ISIS vs. far right extremists, Berger, J.M. 2016. *Nazis vs. ISIS on Twitter* (p. 32). The Centre for Extremism at George Washington University. The broad consensus around 'terrorism' which reigned around the heyday of ISIS would soon fade, however: in 2019 with the Christchurch attack introducing the difficulty of dealing with the seldom designated, and less centralised, violent far right; and later, in 2023, with content relating to Hamas, Palestine and Hezbollah in the post-October 7th war in the Middle East.

> *These are platforms whose core business is "to make life better" (…) [Now] when it comes to terrorism, (…) no one is going to say that politically it's not serious (…). I think one of the explanations for their change is that, in their minds, they can't give the impression of helping terrorism: (…) there are no communication costs* ***vis-à-vis*** *their public on this. They're not going to lose users because they're fighting terrorism.**

As a result, in negotiations with the French and European authorities, the private sector's stance on terrorist content evolved over time, in response to the various scandals that affected it, and the balance gradually shifted in favour of increasing voluntary cooperation. This was emphasised by a respondent from the German Interior Ministry, when they were asked to characterise the Ministry's relationship with Google, Facebook and Twitter during an interview that took place in 2019:

> *I would characterize the relationship as cooperative, all in all. That said, this is a current assessment from 2019, it would look different if you'd asked me three years ago. Companies have become much more willing to cooperate in this area, in the beginning, we experienced a lot of pushback.*

French respondents clearly experienced a similar shift:

> *The tide has turned, there is much more public pressure now on these issues, which works in our favour, because they [the platforms] are really keen*

> *to position themselves as proactive on these issues. Which wasn't at all the case (...) when we started having this dialogue with them [in 2017], when (...) really the platforms' discourse was to say: "The fight against radicalisation is the governments' problem, so, if your citizens become radicalised, it's because you're not doing your job properly, or because something isn't working in your state. (...) The fight against radicalisation is the responsibility of governments, so we don't feel we have to police our platforms."**

The global context of 'techlash' was perceived by the French authorities as an opportunity to obtain the cooperation of social media corporations on a whole range of issues, starting with terrorist content. Also contributing to this window of opportunity was Donald Trump's presidency, with the changes it brought to the relationship between the companies and the US federal government. While the shift in presidents from Obama to Trump had little actual impact on the official US position – which still vehemently opposed any regulation of their economic giants by the European Union[74] – it did create a certain distance, real or perceived, between major Silicon Valley corporations and the new Republican administration:

> *The platforms will work with us when they feel we're a better ally than the U.S. government (...). It's something [a possibility] that didn't exist under Obama, because there was a really strong fusion,*

74 See for instance Trump's threat to sanction the French wine industry in reaction to the possibility of a tax on Big Tech. Tankersley, J. and A. Swanson. 2 December 2019. "French Wine Could Face 100% Tariffs as Trump Confronts France Over Tech Taxes". *The New York Times*. www.nytimes.com/2019/12/02/business/trump-tariff-france.html.

> *which was a bit shaken by the Snowden affair, but then it came back, and now with Trump, it's great because there's a clash of cultures.**

The French and European authorities' relationships with Facebook, Google and Twitter regarding terrorist content were thus facilitated both by a restricted framing of the subject, which the public authorities strategically crafted as a matter of sovereignty, and by the global context of techlash, which favoured increasing the accountability of social media giants.

STRATEGIC CASTING CHOICES

Third, to progress in their negotiations, public and private stakeholders also gradually adopted a common language. This was facilitated by the sociological profile of the negotiators, as both sectors opted to put forward individuals with a strong cultural and/or professional understanding of both the other party and the subject matter.

More common in the United States, the so-called 'revolving door' circulation between the public sector and the tech industry is also a growing trend in Europe, as exemplified most visibly by former UK deputy prime minister Nick Clegg joining Facebook in 2018 as Head of Global Affairs and Communications.[75] On the platform side, Public Policy and Government Relations departments are often populated by individuals with public service experience who are familiar with the ways and constraints of this environment.[76] Meanwhile, Trust and Safety terrorism specialists often have a background in academia, law enforcement or both – a human resources effort

75 On revolving doors, see Tréguer, F. 2019. "Seeing like Big Tech: Security Assemblages, Technology, and the Future of State Bureaucracy." In *Data Politics: Worlds, Subjects, Rights*, edited by D. Bigo, E. Isin and E. Ruppert. Routledge, pp. 36–48; Thibout. 2021. "Google et l'État fédéral étatsunien."

76 Many civil servants who had worked for the Obama administration, in fact, went on to work for Silicon Valley companies when the Trump administration took over.

which Stuart Macdonald and Andrew Staniforth note is appreciated by their public-sector counterparts.[77] This conscious choice on the part of the companies studied reflects their efforts both to respond to public pressure and to define their role in counter-terrorism, a task that initially felt beyond the mandate of even the largest social media corporations. Facebook's former Chief Security Officer Alex Stamos recalls the internal discussions that led the company to hire Brian Fishman, a former director of WestPoint's Center for Counter Terrorism (CTC), to head the company's counter-terrorism efforts in 2016:

> *When I got to Facebook, the biggest content moderation safety issue was ISIS. (...) And the pressure on the companies was a fascinating thing (...) I'm in a meeting and one of our executives had just come back from the UK, where that executive had been yelled at by David Cameron because horrible things had happened there, that the UK government was blaming on Facebook. (...) And my boss at the time, who was the General Counsel, Colin Stretch, said: "Well, what's our goal here? Let's define, what are the success criteria for us." And the executive who had just gotten yelled at, reasonably, based upon their experience, says: "to defeat terrorism". And, Colin, to his credit, says: "Wow, let's pump the brakes there, we are a private social media company. Is that an appropriate goal for a private company, to defeat terrorism? Like, are we going to do drone strikes? What does that mean?" So, coming out of that, we had started*

77 Macdonald, S. and A. Staniforth. 2023. Tackling Online Terrorist Content Together, p. 19.

> *to sharpen up a little bit – well the goal is that you should not allow terrorists to benefit from what you build, which is a much more reasonable thing for a private company to say. But also that discussion led to hiring Brian, and some other folks like him, because it also demonstrated that there is a real gap in understanding: why are terrorists on these platforms, and what benefit they get [out of it]. (…) It was a weird time for, kind of, what is our responsibility to the world?*[78]

Similar efforts to send in people who knew how to 'speak to" tech companies can also be observed on the public-sector side. The French government's appointment of David Martinon as the Ambassador for Digital Affairs and chief negotiator on the issue of terrorist content is one such example. Indeed, his background combined a certain familiarity with the Silicon Valley ecosystem, acquired during his time as the French consul in Los Angeles, with both knowledge of tech policy and substantial experience in negotiating with private actors, acquired through previous missions in the field of Internet governance. He reflects on the atypical nature of the mission entrusted to him as a French diplomat:

> *I had already been working on international Internet governance for four years, and there we very often negotiated with companies, or at least with private bodies. (…) I was used to this atmosphere, which is indeed unusual for diplomats.**

78 Fishman, B., A. Stamos and E. Douek. 16 October 2023. MC 10/16: Facebook's Ex-Counterterrorism Lead on Moderating Terrorism. Moderated Content podcast. https://law.stanford.edu/podcasts/mc-10–16-facebooks-ex-counterterrorism-lead-on-moderating-terrorism.

These casting choices facilitated day-to-day operations within the various innovative public-private forums for counter-terrorism online that had been set up to establish an iterative dialogue between the authorities and social media giants. Many respondents from the public sector, however, expressed frustration at having to negotiate with local offices, pointing to the opaque and strongly hierarchical internal organisation of the firms studied here, whose policy and decision-making centres are firmly located at their Californian or EMEA headquarters,[79] while local offices play a representative role only, with limited decision-making powers.[80]

INSTITUTIONAL INNOVATION

Launched in the aftermath of the terrorist attacks in Paris in 2015, the EU Internet Forum (EUIF) and Europol's Internet Referral Unit (EU IRU),[81] as institutional innovations, constitute the fourth factor in the establishment of state-firm diplomacy on TVEC in the European Union.

The EUIF became the privileged arena for iterative, high-level public-private dialogue on the issue of terrorist content as early as 2015, paving the way for the TERREG, which was described by one respondent as a "natural outcome". According to various interviews conducted for this research, the EUIF provided a venue for the negotiations that led Microsoft, Facebook, Google and Twitter to launch the SIHD and, later, in 2017, the GIFCT. In their early joint communications, GIFCT firms consistently highlighted the EUIF's

79 Generally located in London or Dublin.

80 Ironically, not unlike the organisation of foreign ministries and their embassies.

81 Vieth, K. 2019. "Policing 'Online Radicalisation'."

role and support, alongside that of the UN CTED.[82] Beyond pushing for industry cooperation, the EUIF was also reportedly key in getting the companies to gather and publish specific transparency data on TVEC, which according to one respondent was a significant "get", given the amount of work it involved:

> *It was in fact because of the EU Internet Forum that some tech companies started calculating terrorism-specific data in the metrics they gave in their Transparency Reports. The EU Internet Forum asked the tech platforms involved for metrics on terrorism takedowns. Before that, metrics were very broad in harm type. So that was a really big one – before that, this specificity of terrorism-related data was not collected. It's very hard to calculate that sort of data, because it means you have to create removal labels for a piece of content that include the granularity of saying: "this is being removed for terrorism", or dangerous orgs, or violent extremism.*

Beyond those tangible outcomes, however, the EUIF's main contribution in the eyes of respondents may have been the working relationships it fostered. When asked about major successes

82 E.g. Facebook, Microsoft, Twitter and YouTube. 31 July 2017. "Global Internet Forum to Counter Terrorism to Hold First Meeting in San Francisco." Facebook Newsroom (blog). https://web.archive.org/web/20200204204550/https://about.fb.com/news/2017/07/global-internet-forum-to-counter-terrorism-to-hold-first-meeting-in-san-francisco; Facebook, Microsoft, Twitter and YouTube. 18 June 2018. "Global Internet Forum to Counter Terrorism: An Update on Our Efforts to Use Technology, Support Smaller Companies and Fund Research to Fight Terrorism Online." GIFCT. https://web.archive.org/web/20220523024947 https://gifct.org/2018/06/18/global-internet-forum-to-counter-terrorism-an-update-on-our-efforts-to-use-technology-support-smaller-companies-and-fund-research-to-fight-terrorism-online.

of the EUIF, a respondent from the European Commission's Directorate-General for Migration and Home Affairs answered that the institution's very existence was a considerable achievement:

> *2015 is long ago, so it seems basic to even mention it now, but I realised the other day, working with countries from different regions, that it really is a platform for public-private cooperation where there are constant relationships between the platforms, the Commission and law enforcement, and where there is a general commitment from all of them to change something. I think this, in and of itself, is a success, because we do not necessarily see this on different topics. I would say the EU Internet Forum really is a success in and of itself.*

A civil society contributor involved with the EUIF since its early days concurred:

> *It's been very much – and I don't mean this in a negative way, I mean it in quite a positive way – like a talking shop. (…) it was an opportunity for the Commission, the Member States and the technology companies to feel each other out on the major issues around – especially in the early days, even in the early years – Islamic State activity and what might be appropriate responses. (…) I think all or most parties were suspicious of each other to begin with, suspicious of each other's motives, and they probably had a right to be. (…) I do think that you do have to meet with, talk with other people in order to build trust, and figure out if you can even build trust,*

> *right? But I do think that has occurred over time. I do think all the various actors in this process realise that there were definite positives to be had from these kinds of discussions.*

In addition to the Forum, which is used for high-level dialogue, the EU's law enforcement agency Europol and its Internet Referral Unit (IRU) also contribute to a form of proximity on an operational level. A private-sector respondent pointed to the usefulness of IRUs for companies, saying that the units are an important component of the cooperation on terrorist content that was established between their company and the EU:

> *We have very close working relationships in Europe, and frankly I credit Europe with a lot of the thought leadership in how companies have organised around this effort. (…) So Europe is, well it's not unusual in this space, they're very active, very engaged (…) They're actually quite cutting-edge, kind of [involved a lot in] our cutting-edge developments I mean to say. (…) Also, Europe has something which the American government doesn't have through these Internet Referral Units, the IRUs. These are parts of the government that take as their State responsibility identifying terrorist content on platforms and referring it to companies. There isn't that sort of that equivalent arm of the US government.*

On both sides of the public-private boundary, therefore, institutional innovations set up to counter terrorism online in the EU were recognised as instrumental to the establishment of trust between the parties. Each side highlighted learning processes they had

gone through, which had helped them reach satisfactory levels of cooperation, e.g. on the standardisation of withdrawal notices into formats acceptable to the firms (public sector), or on the adaptation of content rules to European and French speech norms (private sector).

CONCLUSION

THIS REPORT HAS analysed lived experiences of early negotiations on the division of political labour during the early stages of the governance regime for online terrorist content which emerged in the EU between 2015 and 2019. Despite the cultural fault lines that affect how public authorities and companies conceive of their respective roles in this policy area, all the stakeholders involved saw the ability of platforms to develop and deploy automated moderation technologies as a considerable added value in the fight against terrorism. These capabilities spurred the development of a particular form of diplomacy between the public authorities and large, mainstream social media companies, diplomacy that resulted in cooperative working relationships which, in turn, led to a "co-production" of counter-terrorism online at the EU level.[83]

These negotiations have had profound consequences, reaching far beyond their original protagonists and scope. They extend beyond the issue of terrorist content, as the various public-private discussion *forums* originally set up to tackle this content specifically are now being repurposed to address other types of content deemed undesirable.[84] Also, *measures* developed to counter TVEC are now used as a blueprint for handling other undesirable content. For instance, former French Secretary of State for Digital Affairs Jean-Noël Barrot has advocated vocally for the creation of an organisation, modelled on GIFCT, that would be dedicated to handling disinformation. It remains to be seen whether he will pursue this idea now that he has been appointed foreign minister. Secondly, the regime initially developed around TVEC now extends beyond the major platform companies that actively forged it. While Meta,

83 Bellanova, R. and M. de Goede. 2022. "Co-Producing Security."

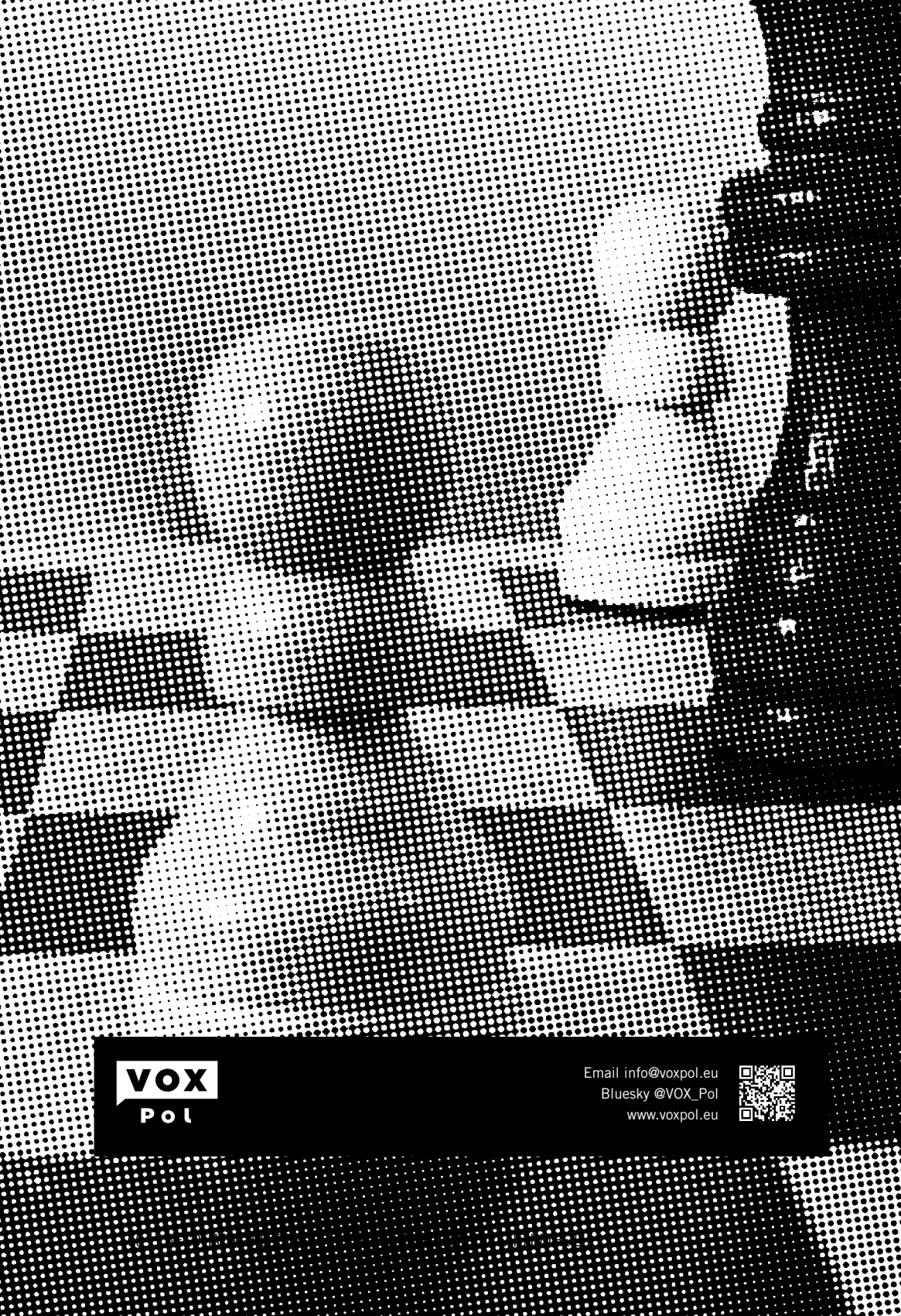
84 In France, the Groupe de Contact Permanent between the French authorities and industry (established after the *Charlie Hebdo* and Hyper Cacher attacks), for example, was redeployed in 2020 following the murder of Samuel Paty, and its mandate was extended to include online hate speech. At the EU level, the EU Internet Forum's mandate was also broadened in 2019, to include CSAM, and again in 2022, to include content relating to drug and human trafficking. Meanwhile, before the EU IRU even began its operations, its scope was expanded to include content relating to human trafficking/illegal migration.

Google and Twitter were the companies that initially attracted most attention from the public authorities, the TERREG applies not just to them but to the entire industry. It formalises their pre-existing informal cooperation with the EU authorities, and imposes it on other companies, which may lack the knowledge, the means or the willingness to comply.[85] Likewise, in a form of "content cartel creep",[86] the 'best practices' established by the largest social media companies within GIFCT are intended to extend to as many firms as possible, including through a now-codified onboarding process, in the form of the Tech against Terrorism mentorship programme.

Looking back on those early negotiations on terrorist content also makes it clear how, in some respects, the dialogue between large platforms and the EU authorities laid the groundwork for more ambitious legislation to come. Early on, public authorities and civil society stakeholders had already identified platform affordances and echo chambers as drivers of radicalisation and polarisation, but they purposely left the issue aside for later, in the face of corporations' firm red line on it. In this respect, the entry into force of the DSA (Regulation EU 2022/2065) in 2022 represents a significant development. While questions remain about how strictly the DSA and TERREG will be enforced, the European authorities' choice of regulation appears validated, in light of recent events. Elon Musk's takeover of Twitter, along with recent layoffs from Google and Meta's Trust and Safety teams, are certainly testing the sustainability of the 'voluntary' elements of the global TVEC governance ecosystem, such as the Christchurch Call and GIFCT.

85 Watkin, A.-L. 2023. "Developing a Responsive Regulatory Approach to Online Terrorist Content on Tech Platforms." *Studies in Conflict & Terrorism*, pp. 1–22. https://doi.org/10.1080/1057610X.2023.2222891.

86 Douek, E. 2020. "The Rise of Content Cartels." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3572309.



VOX Pol

Email info@voxpol.eu
Bluesky @VOX_Pol
www.voxpol.eu