# Co-Producing Security: Platform Content Moderation and European Security Integration

Bellanova, R.; de Goede, M.

DOI
10.1111/jcms.13306

Publication date
2021

Document Version
Final published version

Published in
Journal of Common Market Studies

JCMS Journal of Common Market Studies UACES

JCMS 2021 pp. 1–19 DOI: 10.1111/jcms.13306

# Co-Producing Security: Platform Content Moderation and European Security Integration

ROCCO BELLANOVA[1,2] and MARIEKE DE GOEDE[3*]

[1]Department of Media Studies, University of Amsterdam, Amsterdam, The Netherlands [2]Institute for European Studies, Université Saint-Louis – Bruxelles, Brussels, Belgium [3]Amsterdam Institute for Social Science Research, University of Amsterdam, Amsterdam, The Netherlands

## Abstract

The European Union (EU) seeks to play a leading role in steering the private work of online content moderation, as demonstrated by numerous policy and legislative initiatives in the domain. Two initiatives, in particular, are shaping terrorist content moderation: the creation of a EU Internet Referral Unit and the adoption of a Regulation on preventing the online dissemination of terrorist content (TERREG). This article analyses these initiatives and their practical effects. In particular, it unpacks the *legal* and *technological mechanisms* at the core of EU regulation in the realm of online terrorist content moderation, and how they *co-produce* security decisions across public and private spheres. Based on interviews, fieldwork observations and document analysis, we show how processes of referral and removal, and processes of flagging and filtering are key to EU-directed content moderation. In conclusion, we reflect on content moderation as a novel form of European security integration.

**Keywords:** Europol; IT platforms; security technologies; TERREG; online content moderation; European security integration

## Introduction: Removing Online Terrorist Content

When European Council President Charles Michel visited Vienna in the wake of the November 2020 terrorist attack which left four people dead, he emphasized the importance of developing laws and tools to 'remov[e] terrorist content as quickly as possible when it appears on internet platforms' (European Council, 2020). Michel mentioned this as an important area of European Union (EU)-led counterterrorism, alluding to the then Commission proposal for a Regulation 'on preventing the dissemination of terrorist content online' (EC, 2018b). This Regulation, commonly known as TERREG, has been adopted in April 2021 and seeks to strengthen and steer how social media companies practice *content moderation*.[1] Many social media companies already police what users share on their platforms, to identify and delete harmful and illegal content, including pornography, terrorist-related and copyright-protected material, and content deemed harmful to platforms' brand (Roberts, 2019). Online space is profoundly shaped by content moderation, which is 'essential, constitutional, definitional' to the work of digital platforms, yet normally remains invisible (Gillespie, 2018a, p. 21).

Through TERREG and other initiatives, the EU seeks to play an active role in steering and influencing private practices and decisions on content removal. Content moderation

*Authors' note*: Both authors contributed equally to the article.

[1]Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online [...] OJ L172, 17.5.2021, pp. 79–109. Hereinafter: TERREG.

never happens in a legal vacuum and evolves through confrontation and cooperation between private companies and public authorities in an international context (Langvardt, 2017). The Christchurch Call (2019) – in the wake of New Zealand mosque shootings – was a defining global moment in committing public actors and tech companies to 'eliminate terrorist and violent extremist content online', and set in motion a plethora of initiatives (see also Douek, 2020). In this complex global governance landscape, European authorities have been 'publicly pressing platforms to block certain types of content more aggressively and effectively' than other governments (Cohen, 2019, p. 134). Furthermore, content moderation has become part of the core vision for European security, with European action concerning platforms and online content frontstaged in the most recent Security Union Strategy (EC, 2020a, pp. 13–14) and Counter-Terrorism Agenda (EC, 2020b). Since 2014, the EU has sought ways to control propaganda from the Islamic State (IS) circulated via social media (EC, 2014). A 2019 Europol Report describes a complex online battle, whereby deletions of terrorist-related content are accompanied by quick and creative repostings 'on a wide array of media and file-sharing sites' (Europol, 2019, p. 39). At the same time however, the United Nations' Special Rapporteurs on human rights expressed concern about 'the overly broad definition of terrorist content in the [TERREG] Proposal that may encompass legitimate expression protected under international human rights law' (Kaye *et al*., 2018, p. 2).

This article analyses two EU initiatives on content moderation and their practical effects, to better understand these novel forms of European security and counter-terrorism cooperation. As Sarah Roberts (2019, pp. 14, 27) argues, 'content moderation is a powerful mechanism of control', which raises questions about 'who makes these decisions, how they are made, and whom they benefit.' Platforms – like YouTube or Twitter, but also smaller online service providers – find themselves in the frontline of fighting terrorism and other security threats. Companies identify, select, search, and interpret suspicious datasets; they monitor, regulate, restrict and expel users and user groups (de Goede, 2018). In particular, this article focuses on two EU initiatives that regulate terrorist content moderation and redefine private–public relation in this domain. First, we focus on the EU Internet Referral Unit (IRU), set up in 2015 and then embedded within the Europol European Counter-Terrorism Centre (Europol, 2015 and 2016). IRU's objective is to 'refer terrorist and violent extremist content to Online Service Providers (OSPs) and to provide support to member states in the context of Internet investigations' (Europol, 2020a, p. 4). Between 2015 and 2019, Europol's IRU indentified and referred 111.355 'pieces of [online] content' to social media companies (Europol, 2020b, p. 92). Second, we examine the TERREG Regulation, which aims to align the process of private content moderation with public security and counter-terrorism agendas. TERREG entails hard-law rules for governing platforms' content moderation, legitimises the adoption of high-tech systems for identifying and removing content, and further institutionalises security cooperation between private platforms and Europol. TERREG will come into force in June 2022. Both IRU and TERREG entail new and still opaque configurations of public–private security collaboration within European security.

This article unpacks the *legal* and *technological mechanisms* at the core of EU regulation in the realm of terrorist content moderation, and asks how they co-produce security decisions across public and private spheres. Both IRU and TERREG involve platforms in what we call the *public–private co-production of European security*, thus further

broadening a governance model whereby companies come to play essential role in European security (Bures and Carrapico, 2017; Oliveira Martins and Jumbert, 2020; Carrapico and Farrand, 2021). First, we ask about the legal mechanisms facilitating content removal, and how they build new cross-European databases of suspicious content. How do IRU and TERREG create public–private regulatory structures for sharing and acting upon removal requests? How do IRU and TERREG produce new legal configurations between EU law on the one hand and private, Terms-of-Service regulations on the other? Second, we ask about the technological mechanisms that enable content removal. We ask how these technologies, ranging from standardized interfaces to new databases and algorithms, work in practice and how they redefine European security integration. At what points do private decisions to ban or remove, intersect with actual law enforcing? Taken together, we show how EU-directed content moderation is shaped through practices of referral, removal, flagging and filtering. These practices foster decentralized modes of EU security integration, that depend heavily on public–private co-production.

## I. Theorizing EU Security and Content Moderation

To explore questions about the laws and technologies of EU-directed content moderation, we bring literatures on European security integration in dialogue with research on platform governance and new media studies. Literatures on European security integration help capture important aspects of the institutional and normative questions at stake in the EU's legal and practical involvement in content moderation. First, these literatures have focused on new modes of public–private cooperation in EU counter-terrorism (Kaunert, 2010; Argomaniz *et al*., 2015; Bures and Carrapico, 2017). They have paid particular attention to how EU security practices depend upon novel types of data-sharing across public and private spheres, whereby commercial data, including airline and financial data, are made relevant to public security (Amoore and de Goede, 2005; Bellanova and Duez, 2012; de Goede, 2012; Amoore, 2013; Suda, 2013; Bigo, 2014; Jeandesboz, 2016). Platforms' growing centrality in our societies necessitates the study of how they become involved in European security as frontline law enforcers.

Second, existing literatures analyse the novel institutional forms of European security. Authors explore how European security integration takes the form of networks, bringing together professionals and authorities across member states and even third countries (Den Boer *et al*., 2008; Bicchi and Carta, 2012; Kaunert *et al*., 2013; Cross, 2019). A growing strand of literature studies the security politics of EU agencies (Neal, 2009; Busuioc and Groenleer, 2013; Paul, 2017). For example, research on Europol and Frontex highlights their capacity to support and coordinate security cooperation among national authorities, spearheading security integration through the use of digital technologies (Csernatoni, 2018; Quintel, 2020). Agencies' information capabilities prove essential to European politics at multiple levels, not only providing support on the ground but also influencing policymaking (Carrapico and Trauner, 2013, p. 357). However, little attention has still been paid to the relation between Justice and Home Affairs agencies and private actors, with the partial exception of cybersecurity and terrorism financing (see Bellanova and de Goede, 2020; Carrapico and Farrand, 2021).

Third, this literature has signaled the normative tensions associated with EU security policy-making. Increasing use of personal data and surveillance technologies poses a

specific challenge to EU policy and identity (Bastos and Curtin, 2020). The EU positions itself as a normative power which supports freedom of expression and free media (Diez, 2013; Manners, 2013), for example through programmes empowering civil society in the neighborhood and globally. Yet TERREG, IRU and other policies enable government-backed restrictions on public online speech after minimal review (van Hoboken, 2019, p. 8). Normative aspects of content moderation include both the protection of fundamental rights and freedom of expression, and the safeguarding of the internet as a critical public space and as open to European business. Although 'terrorists' use of the Internet' was already a focus of European security integration in the 2009 Stockholm Program (Argomaniz, 2014, p. 251), literatures on EU studies have yet to focus on platform content moderation as a security *practice* with particular policy relevance in Europe.

Literatures on platform governance, on the other hand, have analysed the changing nature of power relations between internet companies, public authorities and individuals. These works have focused on the growing role of IT companies and (social media) platforms in shaping the internet, which has become a prime public space (for example Gillespie, 2010, 2018a; Crawford and Schultz, 2014; van Dijck and Poell, 2015; van Dijck *et al*., 2018). These works analyse both 'regulation *by* platforms' and 'regulation *of* platforms' (Gillespie, 2018b, p. 254; Ulbricht and Yeung, 2021). The former concerns how IT companies de facto increasingly regulate social and political life, while the latter is understood as how governments seek to regulate platforms (see also Adamski, 2018; Klonick, 2018; Kuczerawy, 2019; van Hoboken, 2019; Ulbricht and Yeung, 2021). Poignantly, Julie Cohen (2019, p. 5) argues that the entanglements between IT companies and public authorities, or between 'code' and 'law,' are redefining the future of our societies. In a similar vein, authors have teased out the implications of the platform economy for public values, including the protection of personal data, but also democtratic control, safety and fairness (Helberger *et al*., 2018; van Dijck *et al*., 2018). This connects to a wider literature unpacking racial bias in Artificial Intelligence and how the design and use of digital technologies can have far-reaching discriminatory effects (Chun, 2009; Benjamin, 2019).

These literatures have paid attention to how content moderation has become a key anchor through which governments and authorities seek to have 'grip' on platforms. As Robert Gorwa (2019, pp. 863–4) notes, the regulation of content moderation is one of the main 'policy levers' used by public authorities to gain some control over these increasingly powerful actors. Appealing to security threats, in particular terrorism, has given public authorities a way of bringing the transnational operations of social media companies within their regulatory remit. The 2017 German Netzwerkdurchsetzungsgesetz (NetzDG) is currently the most far-reaching public regulation of platforms' content moderation anywhere (McMillan, 2019). These appeals have also triggered novel forms of self-regulation and transnational cooperation among platforms and between them and governments in Europe and beyond (Citron, 2018). For example, leading IT companies launched in 2017 a Global Internet Forum to Counter Terrorism (GIFTC), which provides technologies for content moderation at scale (Douek, 2020, pp. 9–10). This literature invites us to further explore how European institutions shape practices for terrorist content moderation, and how platforms – through legal and technological means – become involved in European security integration.

Literatures on content moderation have analysed the practical work of removal decisions and their effects on public online space. Roberts (2019) was one of the first to unpack how privately generated content (such as Facebook posts or YouTube videos) becomes inscribed with suspicion, and to analyse the interactions between platform users, professional content moderators, and, sometimes, police authorities. Content moderation encompasses much more than (alleged) terrorist materials, it includes also hate-speech, harmful content, copyright infringments, and other categories. Indeed, the removal of 'Child Sexual Abuse Material' (CSAM) preceded the terrorism content debate and has generated new governance frameworks and technologies, like 'digital fingerprinting technology' to share and remove harmful material (Douek, 2020, p. 7). Many categories of removal are far from clear-cut however, and rarely without contestation, for example about proper contextualization and explicit legal base. Within this nebulous landscape, regulating terrorist content is inscribed with a particular urgency, especially in Europe. However, with some exceptions (Helberger *et al*., 2018; De Gregorio, 2019; Gorwa, 2019) literature on content moderation is US-focused, and has yet to grapple with the recent developmens in Europe around TERREG and IRU. In addition, existing literatures say relatively little about the modes of practical collaboration between private moderation and public authority, for example how platforms and law enforcement actually interract.

## II. Beyond Public–Private Cooperation: Co-Production

We enquire into the legal and technological mechanisms between private removal and public policing concretely being established by TERREG and IRU. We build on recent calls for practice-focused perspectives within EU studies, that have drawn attention to 'everyday practices' as 'crucial for the performance of European integration' (Adler-Nissen, 2016, pp. 87–8). As Rebecca Adler-Nissen (2016, p. 87) emphasizes, 'routines and habits [...] are integral to making the EU what it is'. This perspective helps analyse how actors relate, cooperate and clash, 'especially in a context of intense privatization of some spheres of public activities' (Bigo, 2016, p. 73).

We suggest that the legal and technological mechanisms for monitoring and removing terrorist online content generate public–private *co-production* of security decisions. Oliveira Martins and Jumbert (2020, p. 12) draw on Science and Technology Studies to show how, in relation to EU border security, co-production between private expertise and public security takes place when 'security experts' play a key role in defining and imagining technological 'solutions', which 'also gives them the power to define of what the problem at stake is'. The notion of co-production helps to move beyond an understanding of a pure privatisation of security, to suggest instead that 'the public and the private are not two realms that can be analysed apart from each other' (Nolte and Westermeier, 2020, p. 63, see also Berndtsson and Stern, 2011). Co-production draws attention to the complex iterative practices at the intersection between public and private (Lindskov Jacobsen and Monsees, 2018, p. 25). It does not imply that public and private actors collaborate seamlessly, but offers a way of 'interpreting and accounting for complex[ity]' (Jasanoff, 2004, p. 3) when security decisions – of referral, removal, flagging and filtering – are produced at the intersection between public policing and private platforms.

We identify two mechanisms of co-production at play in EU-directed content moderation – the legal and the technological. First, we unpack the legal mechanisms through which content moderation works in the field of counter-terrorism, and show how these foster a deep interconnection between public space and private decision-making. The tensions between European law-makers and private companies over new legislation like TERREG are considerable, especially around new removal and reporting obligations for platforms, and the potential for company fines. Yet TERREG intersects with, and indeed works through, private Terms-of-Service rules in complex ways. In focusing on how legal mechanisms operate in practice, we show how public–private co-production of removal decisions takes shape. Second, we focus on the technological mechanisms through which platforms work with public authorities, and how alerts are shared and acted upon. These technological mechanisms are made up of environments and networks where people interact by sharing alerts and digital objects (that is, pictures, text, videos, sound, and so on). Human-machinic interaction is at the core of these technical environments and networks. While leading companies promote the use of artificial intelligence (AI) for assessment and removal of suspicious content among these user-generated digital objects, technological removal mechanisms are far from fully automated.[2] Humans and machines participate to co-produce decisions on referral and removal across public and private spheres, which is crucial to everyday content moderation in Europe and to the shape of European security integration more broadly.

Our research builds upon discourse analysis, observations and interviews with practioners. On the one hand, we conducted a close reading of legislative proposals and texts, technical and legal reports, and strategic documents released by European institutions and companies between 2015 (creation of IRU) and 2021 (adoption of TERREG). These documents provide precious insights about the legal and technological mechanisms, in particular their design and the forms of public–private collaboration that they envision. In addition, we conducted interviews and observations to go beyond the formal documents, and analyse how public–private collaboration is concretely taking shape in this security domain. Notably, we carried out six in-depth semi-structured interviews between 2018 and 2020 with diverse professionals directly working on content moderation, including law enforcement, platforms and NGO representatives. All interviewees have played a key role in either debates about terrorist content moderation or in shaping its practices. They include respondents from security practice, a realm notoriously secretive and difficult to access for researchers (de Goede *et al*., 2019). Finally, two public events organized in Brussels in February 2019 provided further insight into the sometimes conflicting forms of public–private security co-production in the making. One was hosted and co-organized by the Digital Agenda Intergroup of the European Parliament, and the other by Google – thus showing the level of engagement of key actors. Notably, the observation of these events, involving representatives from leading and small platforms, provided us with a better understanding of the power dynamics at play among diverse private platforms.

[2]Google Vice-President for Trust & Safety, public speech delivered in Brussels on 5 Feb. 2019.

## III. Legal Mechanisms: Referring and Removing

This section enquires into the legal mechanisms facilitating content removal, understood as the ensemble of laws and regulations of IRU and TERREG, and their practical uses, that generate the co-production of public–private removal decisions. We show how processes of *referring* and *removing* are at the heart of these new legal practices.

Online content moderation mostly operates on two different legal grounds. On the one hand, social media companies use their own, private Terms of Service (ToS) as legal grounds of content removal. ToS are little-known (and probably rarely read) private agreements that have become powerful quasi-legal forces in the digital world, and that increasingly adopt the language of constitutionalism (Gillespie, 2018a; Celeste, 2019). While TERREG speaks of 'terms and conditions' (Art. 2(8)), platforms like Facebook or YouTube call them 'community guidelines.' Their importance is evident in the fact that they offer ground for decisions that have, in principle, no territorial limitation nor remedy than those afforded by the company itself. On the other hand, national legislation may already offer the legal basis for ordering platforms to remove online content. In this case, online content is deemed to be illegal irrespective of what ToS say. Yet the legal landscape around national legal or administrative removal orders is complex. They generally have a territorial scope, meaning that the targeted content may remain available to users in other countries. However, there are examples of supra-national courts, notably CJEU, which have ordered the transnational removal of internet content (in right-to-be-forgotten cases).[3] In theory, these orders can be contested by both targeted users and platforms themselves, and redress mechanisms are expected to be made available by public authorities. When removals have a transnational dimension, citizen redress becomes more complex.

IRU and TERREG produce new legal configurations between these private, Terms-of-Service regulations on the one hand, and public EU law on the other, realigning public and private actors. As a platform representative summarizes the legal complexities,

> we look at terrorism in a few ways […] we have [in the ToS] a blanket prohibition against terrorism and terrorist organisations using our platforms […] Then of course in many countries […] there are laws against terrorism and dissemination of terrorist content and the EU [TERREG proposal] […] would impact platforms […] so there can be an overlap.[4]

Importantly, in the TERREG proposal, the notion of companies' *duties of care* functions as a legal mechanism by which public actors try to get a grip on the formulation of ToS. Article 3 of the proposal required '[h]osting service providers [ …] [to] take appropriate, reasonable and proportionate actions […] to protect users from terrorist content'. As the European Parliament's amendements to this provision highlight, duties of care create also a 'general obligation […] to monitor the information [platforms] transmit or store' (EP, 2019, Art. 3(1)(a)). TERREG maintains an obligation for platforms to 'include in [their] terms and conditions and apply provisions to address the misuse of [their] services' (Art. 5(1)), with the ultimate effect of further paving the way for public authorities to intervene in platform practices. This may also imply the adoption of technological

[3]Many thanks to Reviewer 2 for pointing this out.
[4]Interview 4, IT platform, Feb. 2019. For the sake of clarity, we have omitted repetitions, pauses and other conversational markers in the quotes from recorded interviews.

mechanisms (discussed below) able to keep social networks clean, and would create new liabilities for companies failing to limit the misuse of their platforms.

*Referral*

A first legal mechanism underlying security co-production is that of referral, whereby suspicious terrorist-related content is reported to the platform by law enforcement. In the case of referral, public authorities like the European IRU act as if they were just another platform user – they report to the company content that goes against plaforms' own 'community guidelines'. All major platforms have their own content moderation systems, often sub-contracted to specialized companies, to identify and take-down online content that is deemed to be in conflict with their ToS (Roberts, 2019). User-reporting mechanisms are important to the content-moderation ultimately carried out by companies themselves. Companies regularly work with so-called 'trusted flaggers', who have access to specific channels of communication, and whose referrals are considered to be strategically important or statistically more accurate, and thus should be prioritized by companies' content moderators (Gillespie, 2018a, p. 131). For example, IRU sometimes functions as such a trusted flagger. In other cases, platforms may provide little or no information about how to flag content.

Since its establishment in late 2015, IRU structurally intervenes in platforms' own content-moderation workflow by identifying and flagging terrorist content (Europol, 2016). IRU's mandate has been formalized in the 2016 Europol Regulation[5] and mainly focuses on terrorist content, in particular Al-Qaida and Daesh propaganda (Europol, 2020a, p. 6). In practice, IRU sends to companies a referral accompanied by a substantiation of the reasons why specific online content violates their community guidelines. The legal mechanism of referral gives public authorities the possibility to use private ToS tactically, fulfilling their counterterrorist objective without leveraging classical judicial means. From a private company perspective, this is how a platform representative describes IRU's referral mechanism:

> [Europol] would say something like they don't have the mandate or the authority to submit referrals to companies under a specific law, so they don't refer content to us and say this violates the law of a member state or of Europe. [But] they feel that is within their mandate to refer content to us that violates our own guidelines, so on the whole, when they refer content […] what they are referring to are policies and terms of service, the community guidelines. […] Europol does have access to that trusted flagger tool along with other NGOs who are expert in terrorism or it be toward child safety or the like.[6]

Importantly, this means that companies *themselves* remain the ones to make the final decision on content removal – and, as one interviewee notes, 'if someone would contest the decision they would contest to the company [ … ], for the […] client signed up to terms […] and conditions'.[7]

[5]Regulation (EU) 2016/794 of the European Parliament and of the Council of 11 May 2016 on the European Union Agency for Law Enforcement Cooperation […] [Europol Regulation], OJ L135, 24.5.2016, pp. 53–114.
[6]Interview 4, IT platform, Feb. 2019.
[7]Interview 3, law enforcement, Nov. 2018.

The referral mechanism brings together two dimensions of law – private and public – leading to public–private co-production in practice. It offers authorities a vicarious way to intervene in the governance of online content moderation. It does so either by influencing how content moderation is already done by companies, or, in the case of those companies that have not set up content moderation workflows, pushing platforms to delete the content brought to their attention by public authorities. It juridically frames online terrorist content as a misuse of platforms' services, and detrimental to the business of companies themselves. At the same time, this mechanism shields public authorities from citizens' direct contestations. It has been critiziced by the European Fundamental Rights Agency (FRA, 2019, p. 36), which argues that

> [T]he concept of referrals [...] leads to broader questions of transparency, effectiveness and accountability [...] it is not clear what the accountability of public authorities would be in situations where they would initiate a removal of legitimate content via referral.

The Fundamental Rights Agency criticism was directly addressed to the TERREG proposal, which included a provision setting up a dedicated referral procedure (Art. 5). Initially, the TERREG proposal asked platforms to 'put in place operational and technical measures' to assess and remove content that was brough to their attention by Europol and other relevant authorities (Art. 5(2), TERREG proposal). It also insisted on the fact that each company had to 'assess the content identified in the referral against its own terms and conditions' (Art. 5(5), TERREG proposal). While this provision has been eliminated in the adopted legislation, yet TERREG still states that the Regulation does not 'preclud[e] the Member Sates and Europol from using referrals as an instrument to address terrorsit content online' (TERREG, §40).

Finally, what is of crucial importance to IRU's referral process is that it creates the legal mechanism for storing suspect online material within Europol itself. For IRU to intervene in the platforms' own content moderation workflow, Europol staff needs to first assess whether the content to be flagged falls within the juridical remit of the agency. If this is the case, then online content (be it a tweet, a picture, or a video) will be referred to the relevant platform *and* stored in a dedicated Europol data-system, called Check the Web Analysis Project. As an interviewee explains:

> As soon as it matches [the agency's] mandate [IRU] store[s] the information. The referral is not the centre of the process, the referral is a second end product but the main one is the core business of Europol which is supporting the member state into the fight against terrorism and that's why [IRU is] looking into this content and [is] harvesting this content.[8]

In other words, datasets extracted from social media platforms become part of Europol databases and can, under specific conditions, be processed for other purposes. Europol's referral mechanism thus feeds a form of European security integration in which online terrorist content extracted from platforms can be stored and further analysed to facilitate cooperation across member states.[9]

[8]Interview 3, law enforcement, Nov. 2018.
[9]Interview 3, law enforcement, Nov. 2018.

*Removal*

TERREG introduces another crucial legal mechanism for online terrorist content moderation, that is, removal orders – which are outside the remit of the IRU. According to TERREG, a national 'competent authority […] shall have the power to issue a removal order requiring the hosting service provider to remove terrorist content or disable access to it' (Art. 3(1)). This mechanism newly creates an obligation on platforms to act upon targeted pieces of online content, resemling traditional juridical practes. At the same time, the removal mechanism in TERREG partially disentangles removal processes from strictly judicial ones. For instance, 'compentent authorities' are not expected to be only the judicial authorities, and Member States may define one or more public agencies that are, in each country, entitled to produce removal orders (TERREG, §35).

Despite its seeming simplicity, the removal mechanism generates complex and far-reaching forms of public–private security co-production. While TERREG provides a common legal basis for competent authorities to request the removal of online terrorist content (Art. 3(4)(d)), it nevertheless triggers issues concerning the jurisdiction of the issuing public authority, the nature of the order itself, the ability to reach out to a company from a third country, and the responsiveness of the platform. TERREG attempts to solve some of these problems through clarifications about what 'competent authorities' are expected to be and do (Arts. 12–13) or by defining a 'procedure for cross-border removal orders' (Art. 4). In view of facilitating the reach of removal orders, TERREG also sets out obligations for platforms operating in the EU from a third country (Arts. 15, 17). While in case of referral public authorities use platforms' policies and architectures, removal mechanisms pushes platforms to formalize their responsibilities vis-à-vis public authorities. This makes platforms more legible and easier to enrol in security cooperation. However, contrary to the mechanism of referrals, TERREG removals co-produce security decisions that are territorially bounded to the EU (Art. 3(1)). This means that, in principle, if companies adjudicate that the content affected by a European removal order does not pose a conflict with their ToS, they may decide to keep it online outside the EU.

The streamlining of legal mechanisms that facilitate interactions between platforms and national authorities paves the way for a stronger role of Europol. The agency will help coordinate removal and referral requests by 'de-conflicting' them – that is, making sure that taking down certain material is not detrimental to ongoing law enforcement or intelligence operations in other member states.[10] As both EP amendments and interviews with professionals show, an increased use of referrals and removal orders may actually hamper ongoing law enforcement operations, because taking down content of a user under investigation may jeopardize a counterterrorist operation carried out in a different member state. The importance of 'deconflicting' investigations is strengthened by the political emphasis on the need to take down content within 'one hour'. If terrorist material should be removed quickly, then de-confliction should be carried out even faster, and routed through a central hub, that is Europol.

The fact that TERREG does not give Europol the possibility to issue removal orders does not seem to be a preoccupation for the agency. As an interviewee noted, '[Europol's] role is to facilitate, […] to help the coordination, standardisation, real time overview [of]

[10]Interviews 2 and 3, law enforcement, Sept. and Nov. 2018.

what is happening and deconfliction, etcetera.'[11] Security decisions in this context do not only require legislation to force companies to align their content moderation to European and national political agendas. They also pressupose a 'meta-regulation' of the public–private co-production of security, able to stabilize the relation between public and private actors and among diverse national authorities (see TERREG, Art. 14).

## IV. Technological Mechanisms: Flagging and Filtering

This section enquires into the technological mechanisms at work in IRU and TERREG, understood as the human-machinic interactions that co-produce removal decisions. Existing literatures have shown the use of (semi-)automated means for content moderation, notably 'algorithimic moderation systems', that are being developed, used and promoted by Big Tech companies (Douek, 2020; Gorwa *et al*., 2020). We enquire into the technological mechanisms created and proposed through IRU and TERREG in the domain of counter-terrorism, and how these bring together humans and machines in the co-production of public–private security decisions. We show how practices of *flagging* and *filtering* are at the heart of these new technical processes.

### *Flagging*

The first technical mechanism is the use of large-scale algorithmic systems that are essential to carry out the first selection – so-called flagging – of digital objects that may be terrorist content. TERREG (§25) provides the possibility for platforms to use 'automated tools,' 'if they consider this to be appropriate and necessary to effectively address the misuse of their services for the dissemination of terrorist content'. This stipulation can be read as a broad encouragement by European authorities for algorithmic systems. These would mainly be used for two purposes – to prevent the further dissemination (for example by re-uploading) of digital objects that have been already labelled as terrorist content, and to identify new terrorist content. The latter is based on machine learning systems able to sift through the material continuously uploaded, to identify those objects bearing the 'signs' of possible terrorist content. Flagged content would then be prioritized in the workflow of human content moderators. As a private company employee describes the human-machinic interaction in the flagging process:

> When content is flagged […] by our own machine learning systems, it goes to these human reviewers and when it comes to terrorism there's a specialist team […] They know the current trends in terrorist activity and in organisations that may have changed their name or merged or morphed or appeared on the stage. [Then] there is someone doing a quality control check to review the accuracy of their, of their flag […], and if they're not scoring a high enough accuracy rate then you'll take corrective measures, you know, more training or, or beyond that.[12]

A law enforcement official, by comparison, argues that algorithms 'allow [IRU] to be more targeted to single out a noise out of the masses.'[13] Thus, even advanced algorithmic

[11]Interview 3, law enforcement, Nov. 2018.
[12]Interview 4, IT platform, Feb. 2019.
[13]Interview 2, law enforcement, Sept. 2018.

systems are not expected to operate alone, but involve interaction with a content moderator, quality control, or security analyst. However, different scenarios of interactions are possible, bringing about diverse ways of co-producing security decisions, and diverse ways of attributing responsibility of those decisions.

Overall, the deployment of machine learning and other algorithmic systems for terrorist content detection shows the interplay of legal and technical mechanisms. ToS are not written directly into code (Lessig, 2006), but rather used to guide human moderators and algorithms. In this context, templates for flagging and referral can be considered an important technical mechanism that make possible the use of legal mechanisms. With regard to cooperation between private companies and Europol, a respondent notes that 'we have a standard text, I'm not even sure that it goes for all the platforms because for some of them it's their own portal, so you just tick boxes on the portal and they receive the content.'[14]

The use of these technologies confirms insights about the crucial role of templates and interfaces for European integration in the field of security (Walters, 2002). Less studied than algorithms, flagging mechanisms do not only facilitate information sharing, but also create multiple configurations of public–private security co-production (Gillespie, 2018a, pp. 87–97, 131–3). For instance, flagging interfaces are, at least in the case of major platforms, different for users, for copyright holders and for law enforcement authorities:

> So if you were flagging as a user you would go to the [platform] video, […] you would click the three dots [next to a video] and you could flag the content under our guidelines. If you were a rights holder, if you were alleging that under law someone's violated your copyright, you could submit illegal removal notice that way. […] [B]ut when you're talking about something like illegal removal notice for terrorism it will be a court or a government official with the ability to flag under the legal removal route and you would flag under the community guidelines route.[15]

In sum, dedicated interfaces may be at disposal of what a company representative calls 'trusted flaggers', that 'can be individuals, NGOs, and sometimes government agencies' and that are 'much more accurate than the average user submitting flags to us, they bring a certain expertise or accuracy to their flagging.'[16]

To enable flagging technically, Europol has collected the templates used by Online Service Providers. It put them, together with 'links to [companies'] portals,' at disposal of those national public authorities that have joined Europol's SIRIUS project (Europol, 2017). Through this project, Europol also makes available a 'library of the terms and conditions of the 50 largest Online Service Providers' (Europol, 2017). TERREG itself introduces templates for standardizing Removal Orders and companies' responses to such requests (Annexes I–III). The templates and interface are low-tech mechanisms (Bonelli and Ragazzi, 2014). As an interviewee notes:

> [H]onestly it's not rocket science, you give context if requested by the platform, you give the URL, so the Unique Identifier, the dates. If it's language that they cannot understand you translate and, but this is about informing them and having enough information for

[14]Interview 3, law enforcement, Nov. 2018.
[15]Interview 4, IT platform, Feb. 2019.
[16]Interview 4, IT platform, Feb. 2019.

> them to take a decision, so if you just send to URL without any context of course this will never work.[17]

Yet, these mechanisms play a central function in the public–private co-production of security decisions. They pivot on what some major platforms already do when moderating content, thus avoiding disrupting the overall private system of policing content. At the same time, they steer those – mostly smaller – platforms that do not yet have formalized content moderation routines, to adopt one that has been standardized by public authorities. Furthermore, the inscription of templates into EU legislation and technical practices, contributes to further the legitimacy of those public authorities reporting content.

### *Filtering*

The second technical mechanism at work in terrorist content moderation concerns the use of so-called *upload filters*. After much debate, the adopted TERREG does not foresee an obligation to introduce proactive measures against the (re)publication of digital objects that have been identified by public authorities as online terrorist content. However, TERREG still leaves open the possibility for platforms to adopt 'automated tools' for filtering content (§25 and Art. 5(2)(d)). Such filters are technical mechanisms to be deployed by private actors, and not by public authorities or IRU. These filters are semi-automated systems that facilitate the identification and removal of digital objects that are expected to breach companies' ToS.[18] Filters for terrorist content are already used across major platforms. Companies like Google increasingly use machine-learning powered filters, that help to detect terrorist content without the continuous intervention of human analysts.[19] In addition, in 2016 leading companies have established a 'hash sharing consortium' (GIFCT, 2021). They have adopted a shared algorithm that produces a non-representative alphanumerical sequence – an hash – of digital objects to be removed (Gorwa *et al*., 2020). These digital fingerprints are stored in a database shared by leading private companies (for example Facebook, Microsoft, Twitter, etc). Every time users upload content to these platforms, its digital fingerprints are run against those stored in the hash database. If there is a match, the content is not published.

Filters may have far-reaching implications in shaping public spaces, because of their encoded pro-active rationale. They moderate content by preventing the publication on platforms of digital objects that have been previously removed or whose content is adjudicated by machine learning as potentially terrorist-related. Sharing hashes in a common, Big Tech-led database, risks creating cascade errors in content moderation, whereby important decisions are aligned on decisions taken by another company according to their own content moderation practice.[20] Contrary to the human expertise used by IRU,[21] filters seem unable to read 'context', thereby further flattening public space. As critics note, filtering promotes preemptive deletion rather than the collection of evidence.[22] From a public–private co-production perspective, these technological mechanisms open up to novel,

[17]Interview 3, law enforcement, Nov. 2018.
[18]IRU does not use filters, but its crawling of online content is supported by algorithmic systems prioritizing security analysts' work, helping them navigating large volumes of data.
[19]Google Vice-President for Trust & Safety, public speech delivered in Brussels on 5 Feb. 2019.
[20]Interview 6, NGO, Feb. 2020.
[21]Interview 1, law enforcement, Sept. 2018.
[22]Interviews 5 and 6, privacy advocate and NGO, Jan. and Feb. 2020.

specific forms of security cooperation. They broaden the scope of private policing in both global and European security, and they strengthen major platforms' role in defining the technical mechanisms that other companies should eventually adopt. Filters would leave a minor role for public law-enforcement, and this – as Douek (2020, p. 10) argues – 'without increasing [platforms'] public accountability.'

Technological mechanisms do not resolve the governance of content moderation. Despite initiatives such as the 'hash databases' or the promotion of machine learning systems, companies prefer not to be bound by the use of specific technological mechanisms. For instance, Twitter (2016) emphasized the complexity of removal judgements, and said: 'there is no 'magic algorithm' for identifying terrorist content on the internet, so global online platforms are forced to make challenging judgement calls based on very limited information and guidance.' Moreover, when carrying out content moderation, law enforcement authorities have to relate with a panoply of companies. Big Tech and social media platforms are not the only companies providing online space hosting terrorist content.[23] In fact, the European Commission (EC, 2018a, p. 36) notes that 'nearly 70% of Europol referrals in 2018 were sent to hosting service providers which can be considered small or micro enterprises'. These companies have limited capacity to handle referrals or removal orders, let alone to create filters or similar technological mechanisms. From a law enforcement perspective, this may become a major hurdle in the co-production of security decisions, as public authorities generally require their active cooperation to take down content. Moreover, public authorities across Europe are also not all equipped to carry out terrorist content moderation, which is resource intensive.

At the same time, these differences in terms of technological capacities across platforms and public authorities invites the fastening of new forms of public–private security co-production. At global scale, they reinforce the diffusion of certain models of private policing of content, whereby some Big Tech companies offer their filtering products to smaller platforms.[24] At European level, TERREG and IRU promote technological mechanisms developed by Europol, such as Internet Referral Management application (IRMA). This is a 'software tool used by [IRU] to help automate the referral process', from 'identifying online terrorist content' to contacting companies (EDPS, 2019, p. 27). Another example is the provision of means to those actors that are not properly equipped to carry out content moderation. As an interviewee puts it, 'as the regulation is calling for Removal Orders, actionable within one hour, you would need to, logically speaking, have a twenty-four-seven system, to cope with the requirements and Europol could provide that.'[25] For instance, the above-mentioned SIRIUS project centres on counter-terrorism and cybercrime, including IRU. It involves other EU agencies and institutions, and national experts for the judiciary and the law enforcement (Europol, 2020a, pp. 9–10). Furthermore, Europol directly engages with platforms, big and especially small. As an interviewee highlights: '[a]nother […] added value of Europol […] is how [the agency] could support the hosting service provider and [in particular] the smaller one […] EU SMEs, that would need some capabilities to […] improve their resilience against the abuse of terrorist group.'[26] Here, Europol plays a pivotal role in the co-production of security,

[23]Interview 1, law enforcement, Sept. 2018.
[24]Interview 6, NGO, Feb. 2020.
[25]Interview 3, law enforcement, Nov. 2018.
[26]Interview 3, law enforcement, Nov. 2018.

not by imposing a security decision, but by creating technological mechanisms that can be adopted by other public authorities and (some) European companies.

## Conclusions

This article has mapped and analysed new European initiatives that work with and through private platforms to flag, filter and remove suspected terrorist online content. We have unpacked the legal and technical mechanisms at work in IRU and TERREG, to understand how European security decisions are co-produced at the intersection between public and private spheres. Our contribution is articulated on three levels – thematically, empirically and conceptually. Thematically, we have furthered attention to the role of content moderation within European security integration, and to how platforms are both structuring and restructured by their relations with European public authorities, especially Europol. This article thus generates a conversation between EU security literature and research in platform governance and new media studies, to focus on the domain of security as a key site for the digital shaping of European public space.

Empirically, the article provides new material about the making of European security practices, understood as both the policy design of how European security is supposed to be organized and function, and the everyday routines of EU professionals and private firms. Notably, we have shown how mechanisms of referral, removal, flagging and filtering, give shape to public–private security collaboration in Europe in the realm of counter-terrorist content moderation. In these new legal and technical mechanisms, Europol plays a crucial role – not as a centralized power but as decentralized hub of coordination, data collection and deconfliction. We also observe the emergence of new legal mechanisms, like the 'duty of care' in TERREG. This duty of care positions social media platforms as *benevolent* security actors – they are pushed to commit themselves to protect the integrity of their services, in a way that is aligned (or at least not opposed) to the priorities of public authorities. Relatedly, we unpack the development of novel mechanisms for flagging and filtering terrorist content. While they configure platforms as frontline security actors, they also provide to leading IT firms and Europol actual means to shape European security integration in a crucial domain.

Conceptually, we have used and developed the term *co-production* to analyse security integration practices in Europe. We have built on literatures that have called for an understanding of the 'everyday of European integration' (Adler-Nissen, 2016, p. 87). We have pushed the notion of co-production beyond its existing focus on the *imagination* of security problems and solutions (Oliveira Martins and Jumbert, 2020, p. 15), to focus on the actual legal and technical mechanisms of security cooperation. When a removal decision is, ultimately, taken on the basis of private ToS after a public authority referral, it is co-produced. While TERREG is part of EU public law, its provisions work *with* and *through* private platforms. Concrete technical mechanisms digitally connect police authorities and private platforms, in ways which make security decisions increasingly difficult to contest. Co-production, then, is a useful conceptual anchor for the dialogue between new media studies' analysis of platform governance on the one hand, and EU studies' analysis of networked security integration on the other. As initiatives for public–private and private–private content moderation proliferate in the wake of the Christcurch Call, the concept of co-production can help map exactly how data are shared, stored and removed.

Content moderation is vital to the future shape of digital public space. It will remain crucial to study how TERREG is practised once it comes into force in 2022, and how it – and related initiatives – informs transnational referral mechanisms and removal decisions.

#### Acknowledgements

We would like to thank Toni Haastrup and two anonymous reviewers for their constructive comments and suggestions, as well as all interviewees for their time and insights. Many thanks to Marie Irmer for her precious research assistance, and to the FOLLOW research team for providing feedback on an earlier version of this paper.

## Funding

This project has received funding from the European Research Council (ERC) under the European Union's H2020 research and innovation program (research project "FOLLOW: Following the Money from Transaction to Trial", Grant No. ERC-2015-CoG 682317) as well as from the Universiteit van Amsterdam Research Priority Area Global Digital Cultures (research project "Digital Platforms and the Digitisation of Expression and Surveillance").

## Notes on Contributors

Rocco Bellanova is Assistant Professor of Critical Data Studies in the Department of Media Studies at the Universiteit van Amsterdam, and Visiting Professor at the Université Saint-Louis – Bruxelles.

Marieke de Goede is Professor of Politics at the Universiteit van Amsterdam and Academic Director of the Amsterdam Institute for Social Science Research (AISSR).

*Correspondence*:
Rocco Bellanova
University of Amsterdam (The Netherlands)
Department of Media Studies
University of Amsterdam
Turfdraagsterpad 9, 1012XT
Amsterdam
The Netherlands.
email: r.bellanova@uva.nl

## References

Adamski, D. (2018) 'Lost on the Digital Platform'. *Common Market Law Review*, Vol. 55, No. 3, pp. 719–51.

Adler-Nissen, R. (2016) 'Towards a Practice Turn in EU Studies: The Everyday of European Integration'. *JCMS: Journal of Common Market Studies*, Vol. 54, No. 1, pp. 87–103.

Amoore, L. (2013) *The Politics of Possibility* (Durham: Duke University Press).

Amoore, L. and de Goede, M. (2005) 'Governance, Risk and Dataveillance in the War on Terror'. *Crime, Law and Social Change*, Vol. 43, No. 2/3, pp. 149–73.

Argomaniz, J. (2014) 'European Union Responses to Terrorist Use of the Internet'. *Cooperation and Conflict*, Vol. 50, No. 2, pp. 250–68.

Argomaniz, J., Bures, O. and Kaunert, C. (2015) 'A Decade of EU Counter-Terrorism and Intelligence'. *Intelligence and National Security*, Vol. 30, No. 2–3, pp. 191–206.
Bastos, F.B. and Curtin, D. (2020) 'Interoperable Information Sharing and the Five Novel Frontiers of EU Governance'. *European Public Law*, Vol. 26, No. 1, pp. 59–70.
Bellanova, R. and de Goede, M. (2020) 'The Algorithmic Regulation of Security'. *Regulation & Governance*. https://doi.org/10.1111/rego.12338
Bellanova, R. and Duez, D. (2012) 'A Different View on the 'Making' of European Security'. *European Foreign Affairs Review*, Vol. 17, No. 2/1, pp. 109–24.
Benjamin, R. (2019) *Race After Technology* (Cambridge: Polity).
Berndtsson, J. and Stern, M. (2011) 'Private Security and the Public–Private Divide'. *International Political Sociology*, Vol. 5, No. 4, pp. 408–25.
Bicchi, F. and Carta, C. (2012) 'The COREU Network and the Circulation of Information within EU Foreign Policy'. *Journal of European Integration*, Vol. 34, No. 5, pp. 465–84.
Bigo, D. (2014) 'The (in)securitization Practices of the Three Universes of EU Border Control'. *Security Dialogue*, Vol. 45, No. 3, pp. 209–25.
Bigo, D. (2016) 'International Political Sociology. Internal Security as Transnational Power Fields'. In Bossong, R. and Rhinard, M. (eds) *Theorizing Internal Security Cooperation in the European Union* (Oxford: Oxford University Press), pp. 64–85.
Bonelli, L. and Ragazzi, F. (2014) 'Low-tech Security'. *Security Dialogue*, Vol. 45, No. 5, pp. 476–93.
Bures, O. and Carrapico, H. (2017) 'Private Security beyond Private Military and Security Companies'. *Crime, Law and Social Change*, Vol. 67, No. 3, pp. 229–43.
Busuioc, M. and Groenleer, M. (2013) 'Beyond Design: The Evolution of Europol and Eurojust'. *Perspectices on European Politics and Society*, Vol. 14, No. 3, pp. 285–304.
Carrapico, H. and Trauner, F. (2013) 'Europol and its Influence on EU Policy-Making on Organized Crime: Analyzing Governance Dynamics and Opportunities'. *Perspectices on European Politics and Society*, Vol. 14, No. 3, pp. 357–71.
Carrapico, H.F. and Farrand, B. (2021) 'When Trust Fades, Facebook is no Longer a Friend'. *JCMS: Journal of Common Market Studies*, Vol. 59, No. 5, pp. 1160–76. https://doi.org/10.1111/jcms.13175
Celeste, E. (2019) 'Terms of service and bills of rights'. *International Review of Law, Computers & Technology*, Vol. 33, No. 2, pp. 122–38. https://doi.org/10.1080/13600869.2018.1475898
Chun, W.H.K. (2009) 'Introduction: Race and/as Technology; or, How to Do Things to Race'. *Camera Obscura: Feminism, Culture, and Media Studies*, Vol. 24, No. 1, pp. 7–35.
Citron, D.K. (2018) 'Extremist Speech, Compelled Conformity, and Censorship Creep'. *Notre Dame Law Review*, Vol. 93, No. 3, pp. 1035–71.
Cohen, J.E. (2019) *Between Truth and Power* (Oxford: Oxford University Press).
Crawford, K. and Schultz, J. (2014) 'Big Data and Due Process'. *Boston College Law Review*, Vol. 55, No. 1, pp. 93–128.
Cross, M.K.D. (2019) 'Counter-terrorism & the Intelligence Network in Europe'. *International Journal of Law, Crime and Justice*, pp. 1–9.
Csernatoni, R. (2018) 'Constructing the EU's High-Tech Borders: FRONTEX and Dual-Use Drones for Border Management'. *European Security*, Vol. 27, No. 2, pp. 175–200.
de Goede, M. (2012) 'The SWIFT Affair and the Global Politics of European Security'. *Journal of Common Market Studies*, Vol. 50, No. 2, pp. 214–30.
de Goede, M. (2018) 'The Chain of Security'. *Review of International Studies*, Vol. 44, No. 1, pp. 24–42.
de Goede, M., Bosma, E. and Pallister-Wilkins, P. (eds) (2019) *Secrecy and Methods in Security Research* (London: Routledge).

De Gregorio, G. (2019) 'Democratising Online Content Moderation'. *Computer Law and Security Review*, Vol. 36, pp. 1–17.
Den Boer, M., Hillebrand, C. and Nölke, A. (2008) 'Legitimacy under Pressure'. *JCMS: Journal of Common Market Studies*, Vol. 46, No. 1, pp. 101–24.
Diez, T. (2013) 'Normative Power as Hegemony'. *Cooperation and Conflict*, Vol. 48, No. 2, pp. 194–210.
Douek, E. (2020) 'The Rise of Content Cartels'. Knight First Amendment Institute at Columbia University.
EC (2014) *Joint statement Malmström–Alfano on the informal Ministerial dinner with IT companies* (Luxembourg: European Commission).
EC (2018a) *Impact Assessment accompanying TERREG [SWD(2018) 408 final]* (Brussels: European Commission).
EC (2018b) *Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online [TERREG - COM(2018) 640 final]* (Brussels: European Commission).
EC (2020a) *Communication […] on the EU Security Union Strategy* (Brussels: European Commission).
EC (2020b) *A Counter-Terrorism Agenda for the EU* (Brussels: European Commission).
EDPS (2019) *Annual Report 2018* (Brussels: European Data Protection Supervisor).
EP (2019) *European Parliament legislative resolution of 17 April 2019 on TERREG* (Strasbourg: European Parliament).
European Council (2020) *Remarks by President Charles Michel after his meeting with the Chancellor of Austria Sebastian Kurz in Vienna* (Brussels: European Council).
Europol (2015) *Europol's Internet Rererral Unit to Combat Terrorist and Violent Extremist Propaganda' [press-release]* (The Hague: Europol).
Europol (2016) *EU Internet Referral Unit. Year One Report* (The Hague: Europol).
Europol (2017) 'SIRIUS project'. Europol, The Hague, available at https://www.europol.europa.eu/activities-services/sirius-project (last accessed on 07.09.2021).
Europol (2019) *Terrorism Situation and Trend Report* (The Hague: Europol).
Europol (2020a) *EU Internet Referral Unit Transparency Report 2019* (The Hague: Europol).
Europol (2020b) *EU Terrorism Situation and Trend Report 2020* (The Hague: Europol).
FRA (2019) *Proposal for a Regulation on preventing the dissemination of terrorist content online and its fundemtal rights implciations. Opinion of the European Union Agency for Fundamental Rights* (Vienna: FRA).
GIFCT (2021) 'Hash-Sharing Consortium'. Global Internet Forum to Counter Terrorism, available at https://gifct.org/?faqs=what-is-the-hash-sharing-consortium-and-how-does-it-work (last accessed on 07.09.2021).
Gillespie, T. (2010) 'The politics of 'platforms''. *New Media & Society*, Vol. 12, No. 3, pp. 347–64.
Gillespie, T. (2018a) *Custodians of the Internet* (Yale: Yale University Press).
Gillespie, T. (2018b) 'Regulation of and by Platforms'. In Burgess, J., Marwick, A. and Poell, T. (eds) *The Sage Handbook of Social Media* (London: Sage), pp. 254–78.
Gorwa, R. (2019) 'What is Platform Governance?' *Information, Communication & Society*, Vol. 22, No. 6, pp. 854–71.
Gorwa, R., Binns, R. and Katzenbach, C. (2020) 'Algorithmic Content Moderation'. *Big Data & Society*, Vol. 7, No. 1, pp. 1–15.
Helberger, N., Pierson, J. and Poell, T. (2018) 'Governing Online Platforms'. *The Information Society*, Vol. 34, No. 1, pp. 1–14.
Jasanoff, S. (2004) 'The Idiom of Co-production'. In Jasanoff, S. (ed.) *States of Knowledge* (London: Routledge), pp. 1–12.

Jeandesboz, J. (2016) 'Smartening Border Security in the European Union'. *Security Dialogue*, Vol. 47, No. 4, pp. 292–309.
Kaunert, C. (2010) *European Internal Security. Towards supranational governance in the Area of Freedom, Security and Justice* (Manchester: Manchester University Press).
Kaunert, C., Léonard, S. and Occhipinti, J.D. (2013) 'Agency Governance in the European Union's Area of Freedom, Security and Justice'. *Perspectices on European Politics and Society*, Vol. 14, No. 3, pp. 273–84.
Kaye, D., Cannataci, J. and Ní Aoláin, F. (2018) *Letter to the EU [doc. OL OTH 71/2018]* (Geneva: Special Procedures of the UN Human Rights Council).
Klonick, K. (2018) 'The New Governors'. *Harvard Law Review*, Vol. 131, pp. 1598–670.
Kuczerawy, A. (2019) *Intermediary Liability and Freedom of Expression in the EU: From Concepts to Safeguards* (Cambridge: Intersentia).
Langvardt, K. (2017) 'Regulating Online Content Moderation'. *Georgetown Law Journal*, Vol., No. 5, pp. 1353–88.
Lessig, L. (2006) *Code Version 2.0* (New York: Basic Books).
Lindskov Jacobsen, K. and Monsees, L. (2018) 'Co-production'. In Hoijtink, M. and Leese, M. (eds) *Technology and Agency in International Relations* (London: Routledge), pp. 24–41.
Manners, I. (2013) 'European [Security] Union: Bordering and Governing a Secure Europe in a Better World?' *Global Society*, Vol. 27, No. 3, pp. 398–416.
McMillan, I. (2019) 'Enforcement through the Network'. *Chicago Journal of International Law*, Vol. 20, No. 1, pp. 252–90.
Neal, A.W. (2009) 'Securitization and Risk at the EU Border: The Origins of FRONTEX'. *Journal of Common Market Studies*, Vol. 47, No. 2, pp. 333–56.
Nolte, A. and Westermeier, C. (2020) 'Between Public and Private: The Co-production of Infrastructural Security'. *Politikon*, Vol. 47, No. 1, pp. 62–80.
Oliveira Martins, B. and Jumbert, M.G. (2020) 'EU Border Technologies and the Co-production of Security 'Problems' and 'Solutions'. *Journal of Ethnic and Migration Studies*, pp. 1–18.
Paul, R. (2017) 'Harmonisation by Risk Analysis? Frontex and the Risk-Based Governance of European Border Control'. *Journal of European Integration*, Vol. 39, No. 6, pp. 689–706.
Quintel, T. (2020) 'Interoperable Data Exchanges within Different Data Protection Regimes: The Case of Europol and the European Border and Coast Guard Agency'. *European Public Law*, Vol. 26, No. 1, pp. 205–26.
Roberts, S.T. (2019) *Behind the Screen* (Yale: Yale University Press).
Suda, Y. (2013) 'Transatlantic Politics of Data Transfer'. *Journal of Common Market Studies*, Vol. 51, No. 4, pp. 772–88.
Twitter (2016) 'Combating Violent Extremism'. Available at https://blog.twitter.com/2016/combating-violent-extremism
Ulbricht, L. and Yeung, K. (2021) 'Algorithmic Regulation: A Maturing Concept for Investigating Regulation of and through Algorithms'. *Regulation & Governance*, e-pub ahead of print. https://doi.org/10.1111/rego.12437
van Dijck, J. and Poell, T. (2015) 'Social Media and the Transformation of Public Space'. *Social Media + Society*, Vol. 1, No. 2, pp. 1–5.
van Dijck, J., Poell, T. and de Waal, M. (2018) *The Platform Society* (Oxford: Oxford University Press).
van Hoboken, J. (2019) 'The Proposed EU Terrorism Content Regulation'. Transatlantic Working Group Paper.
Walters, W. (2002) 'The Power of Inscription'. *Millenium: Journal of International Studies*, Vol. 31, No. 1, pp. 83–108.