UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

# Platforms, Online Surveillance, and the European Union: The Digital Services Act (DSA)

Fahy, R.F.; Zeybek, B.; Bellanova, R.; Möller, J.E.

**Publication date**
2026
**Document Version**
Final published version
**Published in**
European Papers

Link to publication

**Citation for published version (APA):**
Fahy, R. F., Zeybek, B., Bellanova, R., & Möller, J. E. (2026). Platforms, Online Surveillance, and the European Union: The Digital Services Act (DSA). *European Papers*, *11*(1), 563-594.

**General rights**
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

*UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)*

Download date:12 Sept 2026

*European Papers – A Journal on Law and Integration*
Vol. 11, No. 1, 2026, pp. 563–594

***The Future of Digitalisation in EU Law Enforcement***
***edited by*** **Niovi Vavoula**

# Platforms, Online Surveillance, and the European Union: The Digital Services Act (DSA)

*Ronan Ó Fathaigh*[*], *Bengi Zeybek*[**], *Rocco Bellanova*[***] *and Judith Möller*[****]

TABLE OF CONTENTS: 1. Introduction. – 2. State of the art. – 2.1. European online surveillance between content moderation and digital constitutionalism. – 2.2. Platforms and the shaping of information dissemination. – 3. Platforms, online surveillance, and the DSA. – 3.1. Proliferation of government measures targeting online expression. – 3.2. Proliferation of government surveillance. – 3.3. The Digital Services Act (DSA). – 4. Compatibility with human rights. – 4.1. The right to privacy and government monitoring of online platforms. – 4.2. Freedom of expression and government leveraging of platforms. – 4.3. Targeting expression based on broad categorisation of content as illegal. – 5. Conclusion.

ABSTRACT: Over the past few years, governments have been keen on leveraging the power of platforms to further develop their online surveillance practices. The European Union (EU) is becoming a world-leading regulatory powerhouse in this respect. After having adopted in 2021 a Regulation on the countering of online terrorist content (TERREG), the EU legislature adopted the Digital Services Act (DSA) in 2022, which introduces measures boosting the online surveillance capacities and competences of both platforms and governments. While these initiatives are expected to protect the online environment, they risk imposing new forms of restrictions on freedom of expression and the right to respect for private life. Moreover, closer forms of platform/government cooperation raise multiple

[*] Senior Researcher, Institute for Information Law, Faculty of Law, University of Amsterdam, r.f.fahy@uva.nl.

[**] PhD Researcher, Institute for Information Law, Faculty of Law, University of Amsterdam, b.zeybek@uva.nl.

[***] Research Professor, Interdisciplinary Research Group Law, Science, Technology & Society (LSTS), Faculty of Law and Criminology, Vrije Universiteit Brussel, rocco.bellanova@vub.be.

[****] Professor of Empirical Communication Science, Leibniz Institute for Media Research Hans-Bredow & Universität Hamburg, j.moeller@leibniz-hbi.de.

Research for this article was made possible with a grant from the Global Digital Cultures Research Priority Area at the University of Amsterdam, for the project 'Digital Platforms and the Digitisation of Expression and Surveillance'. The paper was tentatively accepted in October 2023 for the Special Issue; and since then, only minor revisions have been made, with the Special Issue accepted in June 2025.

ISSN 2499-8249
doi: 10.15166/2499-8249/883
www.europeanpapers.eu
(CC BY-NC-ND 4.0)

concerns, notably in light of growing debates about the scope and depth of online surveillance provided by platforms' technological affordances. This article applies a trans-disciplinary perspective, bringing together law, political communication and media studies, to better understand the emergence of novel forms of European digital surveillance. First, the article examines how platforms' algorithmic systems shape (and limit) information dissemination. We then critically analyse how EU initiatives expand platforms' surveillance capacities to foster governments' own digital surveillance practices. We focus on the DSA and those regulatory solutions that create novel forms of surveillance. Third, the article assesses how these measures comply with freedom of expression and the right to respect for private life. We conclude by providing recommendations on how to remedy problematic elements of the role that platforms will increasingly play in the digitisation of government surveillance.

KEYWORDS: platforms – surveillance – human rights law – freedom of expression – privacy – Digital Service Act (DSA).

## 1. Introduction

Over the past few years, governments have been keen on leveraging the power of platforms to further develop their online surveillance practices. The European Union (EU) is progressively becoming a world-leading regulatory powerhouse in this respect: after adopting in 2021 a Regulation on the countering of online terrorist content (TERREG),[1] the EU adopted in 2022 the Digital Services Act (DSA),[2] which introduces measures boosting the online surveillance capacities and competences of both platforms and governments. While these two EU Regulations are expected to protect the online environment, they risk imposing new limitations on freedom of expression and the right to respect for private life as guaranteed under European and international human rights law.[3] Moreover, the Regulations, as well as other EU and national initiatives – such as the EU Internet Forum (an EU-platform collaboration to combat terrorist propaganda)[4] and national legislation in France, Germany or Italy[5] – are fostering and institutionalising closer forms of platform/government cooperation[6] in the

[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 (TERREG).

[2] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[3] See Arts 7 and 11 of the Charter of Fundamental Rights of the European Union (2012); Arts 8 and 10 of the Convention for the Protection of Human Rights and Fundamental Freedoms (1950), 213 UNTS 221; and Arts 17 and 19 of the International Covenant on Civil and Political Rights (1966).

[4] European Commission, 'EU Internet Forum: Bringing Together Governments, Europol and Technology Companies to Counter Terrorist Content and Hate Speech Online' (2015), at ec.europa.eu.

[5] See section 3.1. below.

[6] C Sergio, V Mitsilegas, M Stefan and N Vavoula, 'Towards a Principled Level Playing Field for an Open and Secure Online Environment: Regulation, Enforcement and Oversight of Online Content Moderation in the EU and the United Kingdom' (Centre for European Policy Studies 2022), at cdn.ceps.eu.

field of online content moderation.[7] This phenomenon raises multiple concerns, notably in light of growing debates about the scope and depth of online surveillance provided by platforms' technological affordances.

This article focuses on the DSA to investigate how the law facilitates public and private surveillance practices over content disseminated in online platforms. We focus on the DSA because this Regulation provides overarching obligations for digital services that act as intermediaries in their role of connecting consumers with goods, services, and content, while it also foresees possibilities for the government to exercise control over online content. We take a closer look at the DSA from a trans-disciplinary perspective to understand its fundamental rights and surveillance implications. As we show through the remainder of the article, public-private partnership structures that deal with online harms are not novel to the DSA, some of it has origins in self-regulatory initiatives (e.g. EU Code of Conduct on Countering Illegal Hate Speech Online and trusted flaggers)[8] or in national or EU legal frameworks (e.g. criminal procedural law on the orders to provide information). What the DSA does is to introduce an overarching legal framework to govern some aspects of these practices.

This article explores and unpacks the nexus between platforms, the EU and online surveillance. In doing so, it aims at fostering a nascent scholarly conversation at the intersection of law and social sciences, characterised by research that adopts a critical perspective to better grasp the emergence of novel forms of surveillance in digital societies.[9] We thus adopt a trans-disciplinary perspective, combining law, political communication and new media studies, in order to better understand the emergence of novel forms of digital surveillance in the EU. The article is divided as follows: section 2 examines how platforms' algorithmic systems shape, and limit, information dissemination. Section 3 then critically analyses how EU initiatives expand platforms' surveillance capacities to foster governments' own digital surveillance practices, including with a focus on the DSA and those regulatory solutions that create novel forms of surveillance. Section 4 assesses how these measures comply with freedom of expression and the right to privacy. Section 5 concludes.

[7] T Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018); ST Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press 2019).

[8] For example, those that exist under the EU Code of Conduct on Countering Illegal Hate Speech Online, European Commission, at commission.europa.eu.

[9] R Bellanova and M de Goede, 'Co-Producing Security: Platform Content Moderation and European Security Integration' (2022) 60 *Journal of Common Market Studies* 1316; G De Gregorio, 'Democratising Online Content Moderation: A Constitutional Framework' (2019) 36 *Computer Law & Security Review* 105374; T Gillespie, P Aufderheide, E Carmi, Y Gerrard, R Gorwa, A Matamoros-Fernández, ST Roberts, A Sinnreich, and SM West, 'Expanding the Debate About Content Moderation: Scholarly Research Agendas for the Coming Policy Debates' (2020) 9 *Internet Policy Review* 1; J van Hoboken, 'The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications' (Transatlantic Working Group / Institute for Information Law (IViR) 2019), at www.ivir.nl.

## 2. State of the art

Studying platform-government relations in the domain of European online surveillance means engaging with several and diverse strands of scholarly research. Given the objective of this article, we propose to foreground and briefly discuss three lines of research that more closely relate to the scope of our analysis – literatures on online content moderation and on digital constitutionalism on the one hand, and research on the role of platforms and the shaping of information dissemination on the other hand. Each of these lines of research has a sort of disciplinary kernel, respectively (new) media studies, law and communication science. Yet, these lines both intersect each other and cross(-fertilise) diverse disciplinary domains, thus inviting further trans-disciplinary research.

### 2.1. European online surveillance between content moderation and digital constitutionalism

Academic literature on content moderation is currently blossoming. In the most general terms, content moderation can be defined as 'the detection of, assessment of, and interventions taken on content or behaviour deemed unacceptable by platforms or other information intermediaries, including the rules they impose, the human labour and technologies required, and the institutional mechanisms of adjudication, enforcement, and appeal that support it'.[10]

By understanding content moderation as an essential practice of online platforms, this literature emphasises the fact that platforms are far from being a neutral informational space driven by active users. In other words, most of the literatures focusing on content moderation assumes – implicitly or explicitly – that platforms have become particularly power actors, whose seemingly mundane decisions may have far reaching consequences both in relation to their users and to the capacity of public authorities to exercise a form of control over online environment. Hence, as now classical works in this domain suggest, platforms could be better understood as complex socio-technical environments which are far from being 'politics-free'.[11] Hence, somewhat paradoxically, this literature contributes to re-signify the very term platform, now emphasizing how much this kind of actors are far from neutral, but are rather leveraging various forms of governance, notably vis-à-vis online users' behaviour, but also in relation to social order and public policies.[12] In this vein, literature on content moderation tries to foreground and unpack how different forms of

[10] Gillespie, Aufderheide, Carmi, Gerrard, Gorwa, Matamoros-Fernández, Roberts, Sinnreich, and West (n 13) 2.

[11] T Gillespie, 'The Politics of "Platforms"' (2010) 12 *New Media & Society* 347.

[12] R Gorwa, 'What is Platform Governance?' (2019) 22 *Information, Communication & Society* 854; N Helberger, J Pierson, and T Poell, 'Governing Online Platforms: From Contested to Cooperative Responsibility' (2018) 34 *The Information Society* 1. This is why, in this contribution to the special

regulation are exercised in given socio-technical systems.[13] Notably, some of the most innovative research focuses on the relations of power involving users and platforms,[14] that is, how platforms try to control and shape the behaviour of their users, while – at the same time – users develop tactics to bypass some forms of content moderation. Another strand of research casts a light on the otherwise invisibilised relations between the human workforce and the algorithms carrying out the 'detection', 'assessment' and 'interventions' mentioned above.[15] This research increasingly puts into question approaches seeing algorithms as both autonomous and almighty, and rather invites to better grasp the content moderation logics through a focus on actual practices, so to ultimately track their specific dynamics.[16] Finally, some works explores how content moderation is becoming a field in which IT companies and public authorities do not only attempt to shape each other policies – for example by defining new rules for platforms or by influencing policy-making – but also develop new surveillance tools and practices.[17]

Alongside (new) media studies and sociological approaches on content moderation, there is a growing literature about *digital constitutionalism*.[18] From a more legal perspective, both the critical role of platforms with regard to the dissemination of information, and their seemingly capacity to enforce their decision on social networks used by individuals and even public authorities, trigger questions about the need to preserve individuals' fundamental rights in these novel spaces. In this context, digital constitutionalism can be understood as 'the ideology that adapts the values of contemporary constitutionalism to the digital society'.[19] In a sort of feedback loop, the operations through which human rights and values are inscribed into content moderation participate in a 'process of constitutionalisation in the digital environment', which could affect constitutionalism (and this 'even if norms are not yet

issue, we often refer to online service providers as platforms, since it helps us to highlight the power that these IT firms have.

[13] T Gillespie, 'Regulation of and by Platforms' in J Burgess, A Marwick and T Poell (eds), *The Sage Handbook of Social Media* (SAGE 2018) 254-278.

[14] Y Gerrard, 'Beyond the Hashtag: Circumventing Content Moderation on Social Media' (2018) 20 *New Media & Society* 4492; AV Banchik, 'Disappearing Acts: Content Moderation and Emergent Practices to Preserve at-Risk Human Rights–Related Content' (2021) 23 *New Media & Society* 1527.

[15] Roberts (n 11); R Gorwa, R Binns, and C Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 *Big Data & Society* 1.

[16] V Crosset, and B Dupont, 'Cognitive Assemblages: The Entangled Nature of Algorithmic Content Moderation' (2022) 9 *Big Data & Society* 1.

[17] Bellanova and de Goede (n 13).

[18] See G De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press 2022); G De Gregorio, 'The Rise of Digital Constitutionalism in the European Union' (2021) 19 *International Journal of Constitutional Law* 41; N Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4 *Social Media + Society* 1.

[19] E Celeste, 'Digital Constitutionalism: A New Systematic Theorisation' (2019) 33 *International Review of Law, Computers & Technology* 77.

institutionalised or positivised in the hierarchy of legal sources' and even if the process itself 'is not unitary'.[20] Notably, other scholars have sought to deepen the analysis further with a distinct focus on international and European human rights standards, particularly on how human rights law may be operationalised by platforms in their content moderations decisions and practices.[21]

These literatures are already speaking to each other. Besides the fact that they all highlight the crucial role of platforms in shaping our digital societies, they also participate in better understanding the challenges raised by online surveillance for democracies, and how democratic systems should regulate surveillance. Our contribution aims to join this scholarly conversation, which has strong societal and policy relevance. Notably, our goal is to supplement insights of these literatures by discussing ongoing developments at EU level,[22] and how European legislation creates new opportunities to foster government online surveillance. In doing so, we pursue a trans-disciplinary ethos to better grasp the complexity of the EU landscape of online surveillance, which is characterized not only by a wave of novel legislation but also by the unfolding of public-private partnerships giving way to new surveillance practices. Further, we build upon the literature which seeks to demonstrate how legislation, such as the DSA, can be operationalised consistent with international and European human rights standards.

### 2.2. Platforms and the shaping of information dissemination

Platforms have become an essential source of information for news users. In Europe, 40% of news users report they regularly use Facebook for news, 23% use YouTube, 14% use Instagram to access news and 5% frequently turn to TikTok to inform themselves about the world in 2022.[23] This trend is expected to continue as social media platforms offer convenience and the ability to customise your newsfeed to topics users are interested in. The affordances of platforms help users to manage information overload, which is appreciated by users in Europe and beyond.[24] At the same time, users also report that the technological systems that select and present information to them remain opaque to them.[25]

[20] Ibid. 90.

[21] JP Quintais, N Appelman and R Ó Fathaigh, 'Using Terms and Conditions to apply Fundamental Rights to Content Moderation' (2023) 24 *German Law Journal* 881.

[22] For an overview of similar initiatives at an international level, see Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, 'Human rights implications of the development, use and transfer of new technologies in the context of counter-terrorism and countering and preventing violent extremism', A/HRC/52/39 of 1 March 2023.

[23] Reuters Institute for the Study of Journalism, 'Reuters Digital News Report 2022', at reutersinstitute.politics.ox.ac.uk.

[24] N Thurman, J Moeller, N Helberger and D Trilling, 'My Friends, Editors, Algorithms, and I: Examining Audience Attitudes to News Selection' (2019) 7 *Digital journalism* 447-469.

[25] J Swart, 'Experiencing Algorithms: How Young People Understand, Feel about, and Engage with Algorithmic News Selection on Social Media' (2021) 7 *Social Media + Society* 1.

Platforms use algorithmic systems that shape how information is disseminated to their users. These systems mainly serve two main different goals. First, platforms are designed to prioritise content that is likely to be of interest to users and will engage them on the platform. This is achieved through personalisation algorithms that select and adjust content based on past engagement and content selections, the network of the users, and the behaviour of other, similar users. Second, algorithmic systems are used in combination with human supervision to identify content that is in conflict with their user guidelines, such as hate speech, cyberbullying, or disinformation.

Both of these mechanisms, personalisation and content moderation, have a significant impact on what information is accessible to different users. While personalisation arguably increases the relevance of the information presented, it might also create echo chambers where users are exposed only to information that reinforces their existing beliefs and biases. This can limit the diversity of information that users are exposed to. However, empirical evidence suggesting platforms severely affecting content diversity is nuanced. These effects have been noted for specific user groups on specific platforms, but on the other hand, by and large users are not locked in highly individualised so-called filter bubbles.[26]

While personalisation of information determines how visible content is to users, content moderation affects the boundaries for what type of content is allowed. These two mechanisms intersect, when content moderation measures lead to the de-prioritisation of content (often called shadow banning). While less studied so far in the literature, this latter mechanism is particularly noteworthy from a normative and legal standpoint, because users are often not informed that their content is deprioritised, and to what degree the depriorisation takes effect. Because content moderation relies on algorithms to determine what violates platform standards, it is prone to both over-enforcement, whereby non-harmful content is deprioritised or removed, and under-enforcement, whereby content that breaches standards or even the law continues to circulate.

## 3. Platforms, online surveillance, and the DSA

Before examining how European governments leverage platforms to engage in online surveillance and data-gathering, we first explain how this surveillance is built upon a proliferation of new government measures throughout the EU targeting online expression under the guise of illegal and harmful content. This proliferation is key to understanding the growth of government leveraging of platforms for online surveillance, and is notable not only for how widespread in the EU it has become, but

[26] J Möller, 'Filter Bubbles and Digital Echo Chambers' in H Tumber and S Waisbord (eds), *The Routledge Companion to Media Disinformation and Populism* (Routledge 2021) 92-100; H Mercenier, V Wiard, and M Dufrasne, 'Teens, Social Media, and Fake News. A User's Perspective' in G López-García, D Palau-Sampio, E Campos-Domínguez and P Masip (eds), *Politics of Disinformation: The Influence of Fake News on the Public Sphere* (Wiley 2021).

also the serious level of criticism that is being levelled at these measures by international human rights bodies and civil society organisations.

### 3.1. Proliferation of government measures targeting online expression

It is helpful to begin with an overview of EU member states, with approaches from both large and small member states. First, in France, one of the most notable laws targeting online expression is the 2018 Law on the fight against the manipulation of information,[27] which applies to so-called false information on platforms during election periods. The law provides that during an election period, a court can order an online platform to remove 'inaccurate or misleading allegations or imputations of fact', which may 'alter the sincerity of an upcoming vote', and are 'disseminated deliberately, artificially or automatedly', and on a massive scale.[28] Notably, France has not been alone in adopting laws on false news and false information. And there has been such disquiet among human rights bodies over the recent wave of such laws, that four international special mandates on freedom of expression recently came together to make clear that under international human rights standards: prohibitions on the dissemination of information based on 'vague and ambiguous' provisions, including 'false news' or false information are 'incompatible with international standards', and 'should be abolished'.[29]

France went even further in 2020, when the Law on Countering Online Hatred was adopted.[30] The law imposed obligations on platforms to remove 'manifestly illegal' hate speech and other illegal content within 24 hours of receiving a notice, and incredibly, to remove certain terrorist content and child pornography within one hour.[31] The law was subject to trenchant criticism from civil society organisations, arguing it 'seriously undermines' freedom of expression.[32] Importantly, in June 2020, the French Constitutional Council found the law violated the right to freedom of expression, due to the immense 'difficulties' for platforms to assess what was

[27] Loi 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (*Law 2018-1202 on the fight against the manipulation of information*).

[28] Ibid. Art 1.

[29] United Nations Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe Representative on Freedom of the Media, the Organization of American States Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights Special Rapporteur on Freedom of Expression and Access to Information, Joint declaration on freedom of expression and 'fake news', Disinformation and propaganda of 3 March 2017, FOM.GAL/3/17, para 2(a).

[30] Loi n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet (*Law 2020-766 on combating hate speech online*).

[31] Ibid. Art 1.

[32] See for example, 'France: The Online Hate Speech Law is a Serious Setback for Freedom of Expression' (Article 19, 15 June 2020), at www.article19.org.

'manifestly unlawful' content within such short time-frames, and would only 'encourage' platform to remove reported content, 'whether or not' it is 'clearly illegal'.[33]

Second, Germany has enacted one of the most widely known laws targeting expression on platforms, namely the Network Enforcement Act.[34] The law imposes an obligation on platforms to ensure there is a mechanism for users to report 'unlawful content', which is defined as content criminalised under 22 criminal offences under the German Criminal Code, including insult and defamation. Crucially, an obligation is imposed on platforms to check 'whether the content reported in the complaint is unlawful' and whether access to the content 'must be blocked'.[35] In particular, platforms must remove or block access to content that is 'manifestly unlawful' within 24 hours of receiving the complaint. Notably, the Act does not define 'manifestly unlawful'. Again, similar to the French laws, the German law was subject to strong criticism from civil society organisations, scholars, and indeed, the UN Special Rapporteur on freedom of expression was quite forthright with criticism, stating: 'liability placed upon private companies to remove third party content absent a judicial oversight' was 'not compatible with international human rights law'.[36] The UN Human Rights Committee also criticised the law.[37] Nonetheless, Germany went further in 2021, when it enacted a new Law to combat right-wing extremism and hate crime, which amended the Network Enforcement Act.[38] Notably, the law imposes further obligations on online platforms, in particular an obligation to send certain *suspected* illegal content directly to Germany's Federal Criminal Police Authority when it is reported. Part of the law was found to violate EU law by a lower court in March 2022, on various grounds, including that it violated the country of origin principle under Article 3 of the e-Commerce Directive.[39]

[33] Constitutional Court of France, Decision 2020-801 DC (18 June 2020), ECLI:FR:CC:2020:2020.801.DC.

[34] Netzwerkdurchsetzungsgesetz (*Network Enforcement Act*) 2017.

[35] Ibid. section 3(1).

[36] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc OL DEU 1/2017 (1 June 2017) 4. European Digital Rights (EDRi), 'EU Action Needed: German NetzDG Draft Threatens Freedom of Expression' (23 May 2017), at edri.org. See also, R Zipursky, 'Nuts About NETZ: The Network Enforcement Act and Freedom of Expression' (2019) 42 *Fordham International Law Journal* 1325.

[37] Human Rights Committee, Concluding observations on the seventh periodic report of Germany, CCPR/C/DEU/CO/7, 30 November 2021, para 46, at www.ohchr.org.

[38] Gesetz zur Bekämpfung des Rechtsextremismus und der Hasskriminalität (30 March 2021), see J Bayer, 'Germany: New Law against Right-Wing Extremism and Hate Crime' (Inforrm's Blog, 24 April 2021) at inforrm.org.

[39] See C Etteldorf, 'Cologne Administrative Court: New Network Enforcement Act Provisions breach EU Law' (Institute of European Media Law (IRIS), 2022), at merlin.obs.coe.int 23.

Third, in Italy, a notable measure was introduced in 2018 targeting online expression that again drew considerable criticism from human rights bodies. This occurred when the Italian Ministry of the Interior implemented an online reporting mechanism for so-called fake news in the run-up to elections (Operational protocol for the Fight Against the Diffusion of Fake News through the Web).[40] The protocol introduced a 'red button' reporting mechanism where users could 'indicate the existence of a network of content attributable to fake news', and of considerably note was the involvement of the *Polizia Postale*, a unit of the Italian State Police that investigates cyber-crime, which was tasked with reviewing reports and act accordingly. Notably, the UN Special Rapporteur issued a strongly-word review of the Protocol, noting his concern that the Protocol was 'incompatible with the standards of international human rights law', and 'inconsistent with the criteria of legality, necessity and proportionality' under Article 19 of the ICCPR, which guarantees freedom of expression.[41] This was because, according to the Special Rapporteur, important terms such as 'manifestly unfounded and biased news' were not defined and too vague, while the Italian police would become 'arbiters of truth in the public and political domain'.[42]

Fourth, in Spain, recent amendments to the Spanish Criminal Code specifically targeting online content have also been subject to considerable critique, including from the Council of Europe's Commissioner for Human Rights.[43] Notably, Article 578 of the Criminal Code, which criminalises 'glorification' and 'justification' of terrorism' and now provides for increased penalties when such content is disseminated online; while a court may order online platforms to remove such content.[44] In this regard, five UN special mandates were highly critical of the 'vague' and 'broad' notions of 'glorification' and 'justification' of terrorism, and could allow for the 'misuse in the oversight and removal of information available online'.[45] Indeed, human rights organisations have documented the law's application to the removal of

[40] See UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc. OL ITA 1/2018, Operational protocol for the Fight Against the Diffusion of Fake News through the Web. See also S Verza, 'Tackling Fake News, the Italian Way' Resource Centre on Media Freedom in Europe' (Resource Centre on Media Freedom in Europe, 21 May 2018), at www.rcmediafreedom.eu.

[41] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression UN Doc OL ITA 1/2018 cit. 4.

[42] Ibid.

[43] Council of Europe Commissioner for Human Rights, 'Misuse of Anti-Terror Legislation Threatens Freedom of Expression' (Brussels, 3 March 2023), at www.coe.int.

[44] Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal (Spanish Penal Code) at www.boe.es. See also, 'Spain: Briefing on the Penal Code and Threats to Freedom of Expression' (Article 19, 12 March 2020), at www.article19.org.

[45] UN Special Rapporteur on the rights to freedom of peaceful assembly and of association; UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; UN Special Rapporteur on the protection and promotion of human rights and fundamental freedoms

music videos from Twitter and the jailing of Spanish rappers,[46] with Amnesty International going so far as to argue that freedom of expression was 'under attack' in Spain, where that was now a 'climate of self-censorship'.[47]

Importantly, the proliferation of laws targeting online expression is not limited to the largest EU member states. For example, Austria recently introduced its Communications Platforms Act 2021, which requires certain online platforms to remove certain illegal content that has been reported 'no later than 24 hours after receipt of the report'.[48] Further, and as mentioned above, there has been a raft of legislation across EU member states targeting false information and disinformation online. The Commission singled out Hungary for amending Section 337 of the Criminal Code, to include an offence of 'publishing a statement one knows to be false or with a reckless disregard for its truth or falsity' during the state of emergency.[49] In this regard, scholars have examined the legislation in a number of other EU member applicable to disinformation. For example, in Lithuania, under Article 19 of the Law on the Provision of Information to the Public, it is prohibited to disseminate disinformation, which is defined as 'intentionally disseminated false information'. While in Romania, where under a 2020 Presidential Decree, the communications regulator was empowered to order the removal of online content that 'promotes false news' in relation to Covid-19.[50] Further, there are wide range of EU member states that criminalise publishing 'false news' and 'information',[51] which is important to note for the later section which examines the involvement of law enforcement in online monitoring.

while countering terrorism; UN Special Rapporteur on the human rights of migrants; and UN Special Rapporteur on the situation of human rights defenders, Two legal reform projects undermine the rights of assembly and expression in Spain, 23 February 2015, www.ohchr.org.

[46] See for example V Lloret and J Purcell, 'Angry Words: Rapper's Jailing Exposes Spain's Free Speech Faultlines' (The Guardian, 19 February 2021), at www.theguardian.com.

[47] Amnesty International, 'Spain: Tweet … If you Dare: How Counter-Terrorism Laws Restrict Freedom of Expression in Spain' (2018), at www.amnesty.org.

[48] Bundesgesetz über Maßnahmen zum Schutz der Nutzer auf Kommunikationsplattformen (Kommunikationsplattformen-Gesetz) (Communication Platforms Act), section 3 www.ris.bka.gv.at. See 'Austria: The Draft Communication Platforms' Act Fails to Protect Freedom of Expression' (Article 19, 16 October 2020), at www.article19.org.

[49] évi C. törvény a Büntető Törvénykönyvről (2012) (Hungary Criminal Code), s 337 net.jogtar.hu.

[50] Decree of the President of Romania No 195 of 16 March 2020 instituting the state of emergency on the territory of Romania, at www.presidency.ro.

[51] See RÓ Fathaigh, N Helberger, and N Appelman, 'The Perils of Legally Defining Disinformation' (2021) 10 *Internet Policy Review* 1-15 (detailing various laws on 'false news' and 'false information', which are in force in Austria, Croatia, Cyprus, the Czech Republic, France, Greece, Hungary, Malta, Romania, and Slovakia).

### 3.2. Platforms and the proliferation of government surveillance

The previous section sought to demonstrate the proliferation of government measures seeking to restrict expression online across the EU. Crucially, a major consequence of all these new restrictions on online expression, and the leveraging of platform to remove content, is that there had been a related proliferation of government surveillance, monitoring and information-gathering of expression online. Indeed, a whole new array of national and EU agencies are now engaged in monitoring the online environment in some form. For instance, some EU Member States have set up so-called Internet Referral Units (IRUs), which are generally mandated with tasks ranging from the monitoring of specific kinds of online content, to analysing such content or flagging it to platforms so that the latter can moderate it based on their own terms of service.[52] The EU itself has also established such a IRU, which was launched in 2015,[53] within the EU Agency for Law Enforcement Cooperation (Europol).[54]

First, at EU level, public authorities increasingly frame the circulation of online content as a major security concern. For instance, the 2020 EU Security Union Strategy argues that 'core threats to citizens such as terrorism, extremism or child sexual abuse rely on the digital environment'.[55] The EU perspective is that albeit obligations upon platforms are necessary to secure the online space, companies themselves are also regularly abused by some of their users, that not only engage in activities that are deemed illegal by public authorities, but also disrespect the terms and conditions set up by companies.[56] This discourse prefigures practices in which both public authorities and IT companies are actively engaged in the surveillance of online content. Thus, identifying and quickly removing illegal content would benefit not only European security and protect users, but also preserve the legitimate use of online space.

[52] To our knowledge, there is no publicly available, exhaustive list of European countries having already established an IRU. In R Eghbariah, and A Metwally, 'Informal Governance: Internet Referral Units and the Rise of State Interpretation of Terms of Service' (2021) 23 *Yale Journal of Law & Technology* 542, 569, the authors mention the UK, France, 'Germany, Spain, Austria, Belgium, Italy, the Netherlands, and Switzerland'.

[53] Council of the European Union, 'Outcome of the Council Meeting: 3376th Council Meeting (Justice and Home Affairs)' (Brussels, 2015) 4.

[54] For an overview of the EU IRU's mission and mandate, cf. Europol, 'EU Internet Referral Unit Transparency Report 2021: Terrorist Propaganda Monitoring and Analysis, Referrals and Public-Private Partnerships' (2022).

[55] European Commission, 'Communication on the EU Security Union Strategy', COM/2020/605 final 13.

[56] For instance, in the press release accompanying the launch of the EU Internet Forum in 2015, former Commissioner for Migration, Home Affairs and Citizenship, Dimitris Avramopoulos stated that '[t]errorists are *abusing* the internet to spread their poisonous propaganda […] The voluntary partnership we launch today with the internet industry comes at the right time to address this problem. […] This is a new way to tackle this extremist *abuse* of the internet' (emphasis added): cf. European Commission, 'EU Internet Forum: Bringing Together Governments, Europol and Technology Companies to Counter Terrorist Content and Hate Speech Online' (Press Release, 2015), at ec.europa.eu.

There are currently multiple European security initiatives in the field of online surveillance. For analytical purposes, we can identify four types of measures. The first kind is of a policy nature and aims at steering IT companies governance with regard to the scope of content moderation. The organisation of the EU Internet Forum, discussed further below, is a primary example of this type of measures, as it creates a venue where platforms and public authorities (including European institutions) regularly meet and discuss joint initiatives. The second type kind of measures is more operational, devising means for public authorities to intervene directly regarding certain categories of online content. A concrete example of these initiatives is the setting up of IRUs and their content moderation activities, whereby these public authorities actively monitor online content, identify and report – that is, flag – specific content to companies for them to remove it following their terms of service. The third type of initiatives concerns legislative instruments that permit public authorities to acquire the capacity to intervene on platforms in a binding way, for example by issuing so-called removal orders. Across Europe, this is the case with the national legislations as discussed above, as well as the adoption of EU regulations such as TERREG and DSA (the latter discussed below).[57] A fourth type of EU initiatives focuses on fostering companies' capacity in carrying out content moderation in line with European guidelines and regulations. This last strand of initiatives further the idea that public-private cooperation implies also taking into account the actual means (or lack thereof) of both private and public actors, and thus the need to strengthen their technical and operational capabilities, and their capacity to cooperate. For example, the EU Internet Forum has adopted an EU Crisis Protocol in 2019 'to respond rapidly and in a coordinated manner to the dissemination of terrorist content online in the event of a terrorist attack,' involving companies – including the likes of Twitter or Google – as well as public authorities.[58]

All these initiatives highlight the coming into being of what we can call 'symbiotic' relations between law enforcement and platforms.[59] Surely, they differ on important aspects from each other, for example concerning the actual mandate behind the

[57] See also Proposal for a Regulation COM(2022) 209 final – 2022/0155(COD) of the European Parliament and of the Council of 11 May 2022 laying down rules to prevent and combat child sexual abuse. For a critical discussion, see O van Daalen, 'Fundamental Rights Assessment of the Framework for Detection Orders under the CSAM Proposal' (Institute for Information Law, 2023) www.ivir.nl.

[58] European Commission, 'EU Internet Forum: Commission Takes further Steps to Counter Terrorist Content Online' (Brussels, 7 October 2019), at ec.europa.eu.

[59] R Bellanova, and D Duez, 'A Different View on the "Making" of European Security: The EU Passenger Name Record System as a Socio-Technical Assemblage' (2012) 17 *European Foreign Affairs Review* 109, 122. It is important to note that the entanglement between European (and non-European) law enforcement and the private sector is not limited to the domain of online surveillance, but has been a mainstay across multiple security domains, such as border control, counter-terrorism, anti-money laundering and counter terrorism financing, etc. Among many, see L Amoore and M de Goede, (eds), *Risk and the War on Terror* (Routledge 2008); and O Bures, and H Carrapico, 'Private Security

action of public authorities, or the nature of the relation between private and public actors (legal, technical, operational or political). However, these measures influence each other, and they are shaping a rather complex landscape of online surveillance in the EU, which requires a continuous attention to ensuring respect of fundamental rights. Besides focusing on policy and legislative measures such as the EU Internet Forum, the TERREG and the DSA, the EU Internet Referral Unit (IRU) deserves particular attention. This is not only the most concrete and operational initiative so far but it is also becoming a pivotal actor in further shaping the overall landscape. Established in late 2015, its mandate is clarified in the Europol Regulation adopted in 2016.[60] The goal of this Unit is to scan the web in search of terrorist propaganda published by groups such as ISIS and Al-Qaida, as well as content related to human smuggling. The scope of its online surveillance has further broadened in 2021, so that it now 'include[s] [also] right-wing terrorist and violent extremist content online'.[61] Notably, the IRU focuses on online content that is both within Europol mandate and in conflict with the relevant hosting provider's terms and conditions.[62] In these cases, the IRU refers the online content to the platform – asking for taking it down according to the platform's own policy – *and* it stores the same content on a dedicated Europol system – called 'Check the Web' – where it will be processed according to Europol rules and thus may be made accessible to EU member states' law enforcement authorities in the framework of the Europol mandate.[63] As such, IRU operations work at the interface of several practices – the commercial practices of filtering, law enforcement investigation, and supranational cooperation.[64] Furthermore, the IRU is among those public authorities that spearhead public-private partnerships. For example, it engages in cooperation with small and medium companies as well as with major platforms, in the latter case participating in dedicated working groups of the Global Internet Forum to Counter Terrorism (GIFCT), while also providing technical and analytical expertise about, for example, 'the misuse of algorithmic amplification techniques'.[65]

Beyond Private Military and Security Companies: Exploring Diversity within Private-Public Collaborations and its Consequences for Security Governance' (2017) 67 *Crime, Law and Social Change* 229.

[60] 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) and replacing and repealing Council Decisions 2009/371/JHA, 2009/934/JHA, 2009/935/JHA, 2009/936/JHA and 2009/968/JHA.

[61] Europol (n 58) 13.

[62] R Eghbariah and A Metwally, 'Informal Governance: Internet Referral Units and the Rise of State Interpretation of Terms of Service' (2021) 23 *Yale Journal of Law & Technology* 574-6.

[63] According to Europol documentation, Check the Web 'is an electronic reference library of jihadist terrorist online propaganda' which is '[a]ccessible only to law enforcement' and 'contains structured information on original statements, publications, videos and audios produced by jihadi terrorist groups and their supporters'; Europol (n 58) 8. The same document states that the 'goal [of the Check the Web portal] is to improve the EU Intelligence picture on modus operandi of online terrorist propagandists and online CT challenges in EU M[ember] S[tates] and beyond'.

[64] Bellanova and de Goede (n 13).

[65] Europol (n 58) 10-2.

### 3.3. The Digital Services Act (DSA)

Focusing on the DSA allows us to investigate how European Union law can facilitate public and private surveillance practices over content disseminated in online platforms. Despite not being the only EU legislation focusing on content moderation, the DSA deserves our close attention because it provides overarching obligations for digital services that act as intermediaries in their role of connecting consumers with goods, services, and content; it also foresees possibilities for the government to have control over online content. In other words, its scope and potential reach are unique, even compared to legal instruments such as TERREG or national legislations. Thus, we take a closer look at the DSA and we show how it introduces an overarching legal framework to govern some aspects of emerging and existing surveillance practices at the intersection of public-private – platform-government – relations.

Before delving further into the analysis, a brief overview of the DSA is useful for readers not familiar with it. The DSA entered into force in November 2022 and its rules will be applicable directly in all EU Member States for all regulated entities by February 2024.[66] The DSA sets up rules to ensure responsible and diligent behaviour by providers of intermediary services. It aims to create a safer online environment, facilitate the exercise of fundamental rights and attain a high level of consumer protection It cuts through media law, telecommunications regulations, consumer protection, data protection, intellectual property and criminal law but it does not replace sector specific regulation. Briefly put, its provisions range from intermediary liability rules for handling of third party content,[67] a tiered system of due diligence obligations for intermediary services (stricter obligations for platforms with higher number of users), measures on the regulation of content moderation, including through terms of services of platforms and dispute settlement mechanisms for online platforms,[68] systemic risk assessment[69] and independent audit obligations[70] for Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs),[71] to crisis response mechanisms.

In general, several provisions of the DSA carry the imprints of the aforementioned EU policy and regulatory agenda that considers the moderation of illegal and harmful content as a security concern.[72] The DSA is attractive for governments and

[66] Information on important dates on the enforcement of the Digital Services Act can be found in, European Commission, 'The Digital Services Act package' digital-strategy.ec.europa.eu.

[67] Regulation (EU) 2022/2065 (n 6) Arts 4-10.

[68] For example ibid. Arts 14-24.

[69] Ibid. Art 34.

[70] Ibid. Art 35.

[71] Ibid. Art 33; VLOPs and VLOSEs are online platforms and online search engines which have a number of average monthly active recipients of the service in the Union equal to or higher than 45 million.

[72] See section IV.2. and Regulation (EU) 2021/784 (n 5); Communication COM(2020) 607 final from the Commission of 24 July 2020 on the EU strategy for a more effective fight against child sexual

law enforcement: it instrumentalises states' *knowledge* of allegedly illegal and harmful content[73] and operationalises platform capabilities for crime prevention, security and public safety purposes.[74] Under the DSA, platforms' 'responsible and diligent behaviour'[75] is partly defined by their disposition to partner with Member States in their content moderation practices.[76] Although safeguards against abuse of power by platforms and governments are built into the legal framework (e.g. notification to users when their content is removed and dispute settlement mechanisms), concrete measures to mitigate fundamental rights risks are lacking in some instances. The implementation of the DSA is likely to further cement the leveraging of platforms for surveillance and removal of content. This section unpacks these statements by focusing on a number of provisions of the DSA where government influence over online content is most prominent. These are, Article 22 on trusted flaggers, Article 10 on orders to provide information and notification of suspicions of criminal behaviour, and Articles 36 and 48 on crisis response mechanisms and crisis protocols.

*a) Trusted Flaggers*

In general, trusted flaggers are specific third parties with expertise on dealing with certain illegal or harmful content that notify such content. Their progression can be connected the notice and take-down regimes endorsed in the intermediary liability rules in the EU (e-Commerce Directive) and the US (the Digital Millennium Copyright Act).[77]

The DSA introduces a regulated 'trusted flagger' regime with Article 22,[78] and attaches privileges to this status: the obligation for platforms to ensure that the notices of allegedly illegal content submitted by 'trusted flaggers' under Article 16 DSA's notice and action mechanism are processed and decided upon 'with priority and without delay'.[79]

The trusted flagger status under the DSA is to be awarded by the Digital Services Coordinators (DSC). And it is available to a 'limited'[80] number of entities who are

abuse; Proposal for a Regulation COM(2022) 209 final (n 61); see also, M Borelli, 'Social Media Corporations as Actors of Counter-Terrorism' (8 August 2021) 25 *New Media & Society*, at doi-org.kuleuven.e-bronnen.be.

[73] See Regulation (EU) 2022/2065 (n 6) Arts 22 and 10.

[74] Ibid. Art 18.

[75] Ibid. recital 3.

[76] See for example ibid. Art 18 on Notifications of suspicions of criminal behaviour.

[77] See further S Schwemer, 'Trusted Notifiers, and the Privatisation of Online Enforcement' (2019) 35 *Computer Law & Security Review* 1; and N Appelman and P Leerssen, 'On "Trusted" Flaggers' (12 July 2022) 24 *Yale Journal of Law & Technology* 452.

[78] The practice of trusted flagging has been encouraged under (regulated) self-regulatory frameworks – particularly in relation to content the dissemination of which has been considered 'security matters' as mentioned above such as extremism or child sexual abuse content. See INHOPE, www.inhope.org and recital 22 and Art 5(2)(b) of Regulation EU 2021/784 (n 5).

[79] Regulation (EU) 2022/2065 (n 6) Art 19(1).

[80] Ibid. Art 19(1).

'independent from any platform' and 'have particular expertise and competence in tackling illegal content and work in a diligent, accurate and objective manner'.[81] Entities awarded with 'trusted flagger' status by the DSC will be published by the Commission in a publicly available database.[82]

Internet referral units of national law enforcement authorities, Europol, and other semi-public bodies such as organisations 'committed to notifying illegal racist and xenophobic expressions online' are among the entities that can be awarded trusted flagger status by the DSC.[83] This means notifications of allegedly illegal content, in the absence of a court order by law enforcement authorities, if awarded trusted flagger status, will have to be dealt with without undue delay as per the notice and action mechanism of Article 16 of the DSA.[84] This can incentivise platforms towards arbitrary treatment of content under opaque conditions, ranging from outright removal to downranking or otherwise,[85] especially given the threat of liability attached to notice and action mechanisms under Article 16, and the lack of a concrete definition of 'illegal content'.[86] Cooperation between platforms and the law enforcement was recently highlighted in a letter of the European Commission to Alphabet, Inc., which called on YouTube's provider, Inc., to respond promptly to the requests of Europol and law enforcement authorities given the surge of *alleged* illegal content and disinformation following the conflict between Hamas and Israel.[87] While the letter pushes platforms to take measures to deal with illegal content and disinformation pursuant to their obligations under the DSA and warns against possible penalties for non-compliance, it does not touch upon censorship risks such measures can entail.

As a form of privatised enforcement of online content, accountability, transparency, and legitimacy concerns and risks to fundamental rights are part of flagging arrangements.[88] Such concerns are further crystallised with the formal recognition of government involvement in flagging under the DSA (e.g. the possibility of Europol or Internet Referral Units getting trusted flagger status). In principle, state intervention

[81] Ibid. recital 61.

[82] Ibid. Art 22(5).

[83] Ibid. recital 61.

[84] Ibid. recital 62 states, '[…] Given that trusted flaggers have demonstrated expertise and competence, the processing of notices submitted by trusted flaggers can be expected to be less burdensome and therefore faster compared to notices submitted by other recipients of the service […]'.

[85] See section II.

[86] The Digital Services Act adopts the following definition of 'illegal content': 'any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law' Regulation (EU) 2022/2065 (n 6) Art 3(h).

[87] T Breton, Letter of 13 October 2023, at twitter.com and C Goujard, 'YouTube Gets Reminder to Follow EU Law over Hamas Content' (Politico, 13 October 2023), at www.politico.eu.

[88] Schwemer (n 81).

on speech is subject to a higher level of constitutional scrutiny,[89] but co-opting platforms by public entities to enforce illegal content bypass constitutional safeguards. In this regard, for example, the Council of Europe's Committee of Ministers has explicitly stated that 'State authorities should obtain an order by a judicial authority or other independent administrative authority, whose decisions are subject to judicial review, when demanding intermediaries to restrict access to content'.[90]

Despite the risks of indirect state intervention over online speech in flagging, the DSA does not contain *specific* measures to address accountability and transparency issues to safeguard freedom of expression and the rule of law.[91] The DSCs' competencies when awarding the flagger status is limited to reviewing whether an entity fulfils the conditions to be a trusted flagger in consideration of flaggers' expertise and diligence.[92] The DSA's intervention to mitigate the risks of state involvement to ensure accountability and transparency in flagging is limited. One mechanism that assists with oversight is the publication of the details of entities awarded with trusted flagger status by the European Commission and the reports on the number of notices submitted and notices categorised platforms, the type of content, and the action taken by platforms.[93] Another one is the possibility to revoke to the trusted flagger status if it submits precise, inaccurate or inadequately substantiated notices.[94] But these provisions do not necessarily address fundamental risks and accountability concerns of flagging. For example, whether a flagger takes measures to safeguard fundamental rights such as to prevent censorship of legal but harmful content when submitting notifications or standards for determining the independence of flaggers (e.g. flaggers' funding schemes) are not explicitly part of the DSA's flagger structure. Other provisions of the DSA related to content moderation such as information requirements[95] and dispute settlement mechanisms[96] that would apply to action taken upon notifications submitted by flaggers could assist the scrutiny of state involvement in flagging. But the DSA does not necessarily seek to reveal the conditions of the relationship between governments, including law enforcement authorities, and plat-

[89] Ibid.

[90] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States of 7 March 2018 on the roles and responsibilities of internet intermediaries, section 1.3.2.

[91] Schwemer (n 81); Appelman and Leerssen (n 81).

[92] The Digital Services Coordinator that awarded the trusted flagger status can suspend or revoke 'trusted flagger' status if it finds that a trusted flagger no longer satisfies these conditions, especially if a trusted flagger 'submitted a significant number of insufficiently precise, inaccurate or inadequately substantiated notices', upon an investigation at its own initiative or on the basis of information received by online platforms and third parties. Regulation (EU) 2022/2065 (n 6) Art 22(6) and (7).

[93] Ibid. Recital 62 and see below Section V.

[94] Ibid. Art 22(6) and (7).

[95] For example, information requirements notifications have to fulfil under Regulation (EU) 2022/2065 (n 6) Art 16(2) and statement of reasons under Art 17 (3)(a), (b) and (d).

[96] Dispute resolution options to be made available by online platforms under Arts 20 and 21.

forms. Provisions of the DSA do not address how and under what conditions platforms respond to law enforcement requests or how public flagger requests connect to the overall content moderation regimes of platforms – for example, whether content flagged by public authorities is automatically taken down.[97]

One final point on trusted flaggers is that the DSA institutes 'parallel structures'[98] for flagging – trusted flagger regime does not streamline all flagging practices into a single legal framework. Public-private partnership can still continue outside of the DSA's trusted flagger scheme, with regard to illegal content, as well as content that is deemed to be 'harmful' or that violates platforms' terms and conditions.[99] As such, platforms continue to be the long arm of the states, which risks circumvention of constitutional safeguards for the protection of fundamental rights, in particular, freedom of expression.[100]

*b) Orders to provide information and notification of suspicions of criminal offences*

Another notable provision of the DSA that reinforces public-private partnership in dealing with illegal or harmful content is its provisions that facilitate data gathering from platforms of user information. Article 10 of the DSA allows national judicial authorities and administrative authorities can order online platforms to provide 'specific items of information' about a user. Upon receipt of the order, platforms are to inform the issuing-authority 'without undue delay' of the effect given to the order.[101] Article 10(2) provides minimum conditions for orders to be satisfied by authorities to give rise to the obligation for platforms to *inform* the relevant authorities about the effect given to those orders. The DSA does not provide an obligation to enforce the order itself.[102]

Are there any safeguards to prevent arbitrary request of user information by national judicial or administrative authorities? Issuing authorities provide to platforms reasons as to why the information is sought to platforms and platforms in turn inform users with statement of reasons and redress opportunities.[103]

[97] Appelman and Leerssen (n 81); Civil society organisations, 'Civil Society Letter to the European Parliament on Terrorism Databases' (2019), at cdt.org; Europol (n 58); J Barata, 'Terrorist Content Online and Threats to Freedom of Expression' (Verfassungsblog, 14 March 2022), at verfassungsblog.de.

[98] Appelman and Leerssen (n 81).

[99] Regulation (EU) 2022/2065 (n 6) recital 62; see also, A Kouwenhoven, E Rosenberg, and R van der Poel, 'NCTV volgt heimelijk burgers op sociale media' (NRC, 9 April 2021), at www.nrc.nl; W Crisp and V Dodd, 'Met Police Profiling Children "on a Large Scale", Documents Show' (The Guardian, 3 June 2022), at www.theguardian.com; Facebook Oversight Board, 'UK drill music', 2022-007-IG-MR (22 November 2022), at www.oversightboard.com.

[100] EDRI, 'Human Rights and Privatised Law Enforcement' (25 February 2014), at edri.org; Article 19, 'Regulating Speech by Contract' (Policy brief, 2018), at www.article19.org; B Chang, 'From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU' (2017) 49 *Columbia Human Rights Law Review* 114.

[101] Regulation (EU) 2022/2065 (n 6) Art 10(2).

[102] Ibid. Art 10(2) and recital 31.

[103] Ibid. Arts 10(2)(a)(iv) and 10(5).

But there are exceptions. Article 10(2)(a)(iv) provides that no statement of reasons needs to be given where the reasons relate to the 'prevention, investigation, detection and prosecution of criminal offence'.[104] The breadth of this exception gives considerable room for authorities to order platforms to provide information and data on users without the need to specify the purpose of the order, beyond that it is needed for investigation of a criminal offence. For example, these orders can relate to contextual and fluid content such as content deemed as insult and defamation and/or publishing false news and false information which are criminal offences in a number of EU member states.[105] In cases of exceptions, without prejudice to other EU and national laws, platforms' terms of service define how they deal with the procedural aspects of handling with government orders such as how platforms handle government requests for user information and how users are informed of this are defined by (e.g. notification to users and transparency reports).[106] Voluntary procedural efforts concerning government access to user information are worrying as research suggests that voluntary procedural safeguards fall short of safeguarding user information against unjustified intervention on the right to privacy by governments as the terms and conditions of platforms offer lower fundamental rights safeguards and their enforcement significantly varies between platforms.[107]

Further, under Article 18, where a platform becomes aware of information giving rise to a suspicion that a serious criminal offence has taken place, is taking place or is likely to take place, the platform must 'promptly inform' national law enforcement of its 'suspicions' and 'provide all relevant information available'. Notably, there is no definition of 'all relevant information', and there are no procedural safeguards stipulated under Article 18. Thus, Article 18 basically co-opt platforms into an information-gathering role for law enforcement.

*c) Crisis response mechanisms and crisis protocols*

Perhaps the most striking examples of measures that enhance public and private surveillance capacities in the DSA are Article 36 on crisis response mechanisms and Article 48 on voluntary crisis protocols.[108] These provisions leverage platforms' role in dissemination of content in crisis situations for public safety, security and counter-

[104] Ibid. Art 10(2)(a)(iv).

[105] See European Regulators Group for Audiovisual Media Services (ERGA), 'Report on Notions of Disinformation and Related Concepts' (2020), at media-board.europa.eu.; and Fathaigh, Helberger, and Appelman (n 55).

[106] Google, 'How Google Handles Government Requests for User Information', at policies.google.com; and E Kosta and M Brewczyńska, 'Government Access to User Data: Towards more Meaningful Transparency Reports' in R Ballardini, P Kuoppamäki and O Pitkänen (eds), *Regulating Industrial Internet through IPR, Data Protection and Competition Law* (Kluwer Law International 2019) 253.

[107] Kosta and Brewczyńska (n 110).

[108] See further D Buijs and I Buri, 'The DSA's Crisis Approach: Crisis Response Mechanism and Crisis Protocols' (DSA Observatory, 21 February 2023), at dsa-observatory.eu.

terrorism purposes. Crisis protocols and response mechanisms under DSA are extensions of the EU security policy initiatives,[109] which rely on private and public cooperation. Among the key priority areas of EU security policy is to tackle evolving treats, such as disinformation, and to prevent and counter the spread of illegal content online such as terrorism, extremism or child sexual abuse.[110] Crisis protocols are not new to the DSA and they existed as part of the aforementioned EU security policy. For example, the EU Internet Forum adopted the EU Crisis Protocol in 2019 in the context of terrorist or violent extremist content.[111]

Article 48 on voluntary crisis protocols was in the European Commission's initial proposal of the DSA and the crisis response mechanism in Article 36 was added during the third trialogue in March 2022.[112] The definition of 'crisis' is found in Article 36(2) of the DSA, 'extraordinary circumstances lead to a serious threat to public security or public health in the Union or in significant parts of it'.[113]

According to these provisions, crisis situations may require *certain specific measures* to be taken urgently by providers of very large online platforms.[114] VLOPs or VLOSEs decide the substance of the measures to take but the Commission upon a recommendation by the European Board for Digital Services (the Board) can foresee an obligation to take measures, which afford them considerable influence over content moderation processes of platforms. For example, as given in Recital 91, the Commission can issue orders to intensify 'cooperation with trusted flaggers' such as law enforcement for removal of content, to a specific point of contact for trusted flaggers during a crisis and to adapt their terms and conditions.[115] The Commission can also order platforms to assess how their functioning and use of their services significantly contribute or are likely to do so to serious threats to public safety or public health.[116] In addition, times of crisis may also permit the Commission to initiate voluntary crisis protocols that can concern online platforms and online search

[109] European Commission, 'Communication on the EU Security Union Strategy' COM(2020) 605 final (24 July 2020).

[110] Ibid. and European Commission, *Tackling Evolving Threats*, at commission.europa.eu; see also, European Data Protection Board and European Data Protection Supervisor, the Joint Opinion 04/2022 on the Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse of 28 July 2022.

[111] European Commission (n 62).

[112] L Bertuzzi, 'DSA: European Commission Pitches Crisis Management Mechanism, Supervisory Fees' (Euractiv, 23 March 2022), at www.euractiv.com.

[113] Regulation (EU) 2022/2065 (n 6) Art 36(2) and recital 91; further adds, 'Such crises could result from armed conflicts or acts of terrorism, including emerging conflicts or acts of terrorism, natural disasters such as earthquakes and hurricanes, as well as from pandemics and other serious cross-border threats to public health'.

[114] Ibid. recital 91.

[115] Ibid. recital 91 and Art 36.

[116] Ibid. Art 36(1)(a).

engines in addition to VLOPs and VLOSEs to address crisis situations, 'strictly limited to extraordinary circumstances affecting public security or public health'.[117]

Against this background, Articles 36 on crisis response mechanisms and 48 on crisis protocols further embed law enforcement and unaccountable public intervention on the online environment during a crisis. The implementation of these provisions can lead to a considerably more aggressive private-public enforcement of rules governing content. Crisis protocols and response mechanisms have been criticised as 'an overly broad empowerment of the European Commission to unilaterally declare an EU-wide state of emergency'[118] when they were initially proposed by the Commission. This criticism might still hold as situations of crisis, which lack a consistent and clear definition, grant the Commission, an *executive* body, extensive powers to intervene on fundamental rights in the form of an *administrative act*.

Fuelled by ensuring safety and security online, Articles 36 and 48 continue to appoint platforms to make judgments on the permissible limits of content on behalf of the states, and bypasses constitutional safeguards for government action. Any restriction to freedom of speech must be proportionate, respect the essence of that right, and meet objectives of general interest to the EU pursuant to Article 52(1) of the EU Charter of Fundamental Rights.[119] Strict obligations to ensure proportionality of measures are lacking, as, for example, the Commission is to 'aim to' ensure that the protocol in question sets time limits and defines the specific parameters to define specific extraordinary circumstances when drawing up crisis protocols.[120] It is difficult to imagine how measures taken under Articles 36 and 48 can fulfil necessity and proportionality requirements under the Charter, especially given the broad reach of VLOPs (e.g. the number of users that would be implicated by crisis response mechanisms), that these measures are taken by an executive body in the absence of judicial review and parliamentary control, and the lack of strict and specific restrictions on the Commission's actions.[121]

Times of crisis may require extraordinary measures to be taken. But the legal framework that allows such measures and their implementation must comply with the principles of the rule of law and observe strict proportionality requirements. Some parallels from the Council of Europe legal regime can be drawn in terms of safeguarding fundamental rights in times of crisis. For example, during Covid-19

[117] See ibid. Art 48 and recital 108 gives examples of cases of such extraordinary circumstances: 'where online platforms are misused for the rapid spread of illegal content or disinformation or where the need arises for rapid dissemination of reliable information'.

[118] 'EU: Digital Services Act Crisis Response Mechanism must Honour Human Rights' (Article 19, 13 April 2022), at www.article19.org.

[119] Case C-417/11 *Council v Bamba*, EU:C:2012:718.

[120] See Regulation (EU) 2022/2065 (n 6) Art 48(4).

[121] Art 19 (n 122); Venice Commission of the Council of Europe, 'Rule of Law Checklist' (2016), at www.venice.coe.int.

Hungary approved provisions criminalized 'fake news',[122] which was criticized by the Council of Europe Commissioner for Human Rights as an erosion of media pluralism and freedom of expression in Hungary.[123] Further, the Committee of Ministers of the Council of Europe recommends Member States to avoid using vague terms when imposing restrictions of freedom of expression and information in times of crisis and to clearly define 'incitement to violence and public disorder'.[124] The ECtHR held that emergency measures should go through democratic review in order to ensure that they are 'strictly required by the exigencies of the situation', as laid down in Article 15 of the ECHR, and that they are appropriate responses to the state of emergency.[125] In the context of emergency measures, the ECtHR further highlighted that democracy thrives on freedom of expression and that the existence of a 'public emergency threatening the life of the nation' should not serve as an excuse for undue limitations on freedom of expression.[126]

To conclude, our explanations above demonstrate that a number of provisions of the DSA integrate governments into the functioning and operation of private platforms, the conditions of which are not always transparent. The DSA endorses government efforts to detect and tackle illegal activity online, to prevent societal harms and to maintain public order, public health and public safety during times of crisis without strict fundamental rights safeguards against potential abuse. Further, some provisions of the DSA demonstrate the securitisation of platform regulation which exasperates the outstanding proxy censorship concerns. Historically, fundamental rights, including freedom of expression and the right to privacy have been core to establishing the duties and responsibilities of platforms. The shift towards stronger government involvement in online speech enforcement and intensified surveillance on platforms embodied by the EU approach to platform regulation, as also observed in the DSA, are alarming for the protection of fundamental rights online, including the right to freedom of expression and the right to privacy.

## 4. Compatibility with human rights

Having discussed the DSA and TERREG, setting out how European governments are increasingly leveraging platform to restrict expression online, and the consequent

[122] L Bayer, 'Orbán Critics Fall Foul of Hungary's Controversial Corona Law' (Politico, 14 May 2020), at www.politico.eu; O Marzocchi, 'The Impact of Covid-19 Measures on Democracy, the Rule of Law and Fundamental Rights in the EU' (23 April 2020) *European Parliament Briefing* www.europarl.europa.eu.

[123] Council of Europe Commissioner for Human Rights, 'It is High Time for Hungary to Restore Journalistic and Media Freedoms' (30 March 2021), at www.coe.int.

[124] Guidelines of the Committee of Ministers of the Council of Europe on protecting freedom of expression and information in times of crisis (26 September 2007), at search.coe.int.

[125] *Alparslan Altan v Turkey* App no 12778/17 (ECtHR, 16 April 2019) para 118.

[126] *Dareskizb Ltd v Armenia* App no 61737/08 (ECtHR, 21 September 2021) para 70 and *Mehmet Altan v Turkey* App no 13237/17 (ECtHR, 20 March 2018) para 210.

proliferation of online surveillance and monitoring of expressive activity, this section examines the European and international human rights concerns. Importantly, these measures, such as monitoring and gathering user content on online platforms, and government-platform initiatives to remove content from online platforms, readily implicate both the right to private life, and the right to freedom of expression, guaranteed under the International Covenant on Civil and Political Rights (ICCPR),[127] and the European Convention of Human Rights (ECHR).[128]

### 4.1. The right to privacy and government monitoring of online platforms

The first question that must be posed is whether public authority monitoring of an individual's *public* social media content and activity constitutes an interference with the right to private life. Notably, there is considerable literature on mass surveillance methods deployed by intelligence agencies and the compatibility of these measures with human rights law.[129] However, what we are particularly interested in is the monitoring and gathering of publicly-available information from online platforms, and the content individual posts online by law enforcement agencies, and other national and international authorities. This is a particularly pertinent question, given the range of national and European agencies engaged in these practices, where gathered data and information used to have content removed from platforms, and where there is little oversight or transparency.[130] In other words, was the Dutch Ministry of Defence correct when stating that its surveillance unit only relied on 'public information' online in defence of its actions.[131]

Importantly, under international human rights law, it is quite clear that government monitoring of publicly-available information about an individual, such posts on online platforms or information shared publicly online, can constitute a serious interference with the right to privacy under Article 17 ICCPR.[132] Indeed, the right to privacy extends to information 'that is publicly available', including when government agencies monitor, collect or analyse a person's online posts.[133] This is because

[127] International Covenant on Civil and Political Rights (1966) 999 UNTS 171.

[128] Convention for the Protection of Human Rights and Fundamental Freedoms (n 3).

[129] See B van der Sloot, 'The Quality of Law: How the European Court of Human Rights gradually Became a European Constitutional Court for Privacy Cases' (2020) 11 *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 160, 177; and E Kosta, 'Algorithmic State Surveillance: Challenging the Notion of Agency in Human Rights' (2020) 16 *Regulation & Governance* 212.

[130] Kosta and Brewczyńska (n 110) 253-274.

[131] Ministry of Defence, 'Land Information Manoeuvre Centre helpt Defensie anticiperen' (16 November 2020), at open.overheid.nl.

[132] International Covenant on Civil and Political Rights (1966), Art 17.

[133] UN High Commissioner for Human Rights, 'The Right to Privacy in the Digital Age' UN Doc A/HRC/39/29 (3 August 2018) 6.

'public sharing of information does not render its substance unprotected'.[134] Crucially, this means is that any interference with privacy is only compatible with Article 17 in very limited circumstances, where it is 'prescribed by law', 'strictly and strictly and demonstrably necessary to achieve a legitimate aim', and 'proportionate'.[135] necessary. Indeed, the UN Special Rapporteur on privacy has recently criticised the 'widespread absence of legislation' on online monitoring.[136]

Similarly, under European law, government monitoring of public expressive activity can constitute an interference with a person's right to private life under Article 8 ECHR.[137] Importantly, the ECtHR has considered whether the police violate a person's right to private life when they gather and retain information on a person's public expressive activity that may not be illegal, but is considered harmful. This question was at issue in the recent *Catt v the United Kingdom* judgement, where an anti-war activist discovered he was included on a UK police 'extremism' database, which documented information gathered by police about his presence at perfectly lawful public protests and demonstrations.[138] Following police refusal to delete the information from the database, the European Court ultimately found a violation of Article 8 ECHR. Crucially, the Court wholly rejected the UK government's argument that the interference with privacy was 'very limited', and unanimously held that it makes no difference that monitoring and gathering of information was not from 'covert operations', and the information collected was 'in the public domain',[139] holding that the 'mere storing of information' amounted to an interference with Article 8 ECHR.[140]

Notably, the Court was particularly scathing of the police gathering of information, holding the police action 'did not take into account' the 'heightened level of protection' that data 'revealing a political opinion' enjoys.[141] Indeed, the Court stated its retention must have had a 'chilling effect'.[142] Crucially, the Court held that there was no 'pressing social need' to retain the activist's data, and there was an 'absence of effective safeguards', which was of 'particular concern' to the Court.[143] In a particularly important passage, the Court laid down the fundamental principle that col-

[134] Ibid.

[135] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/23/40, 83.

[136] UN High Commissioner for Human Rights (n 137) para 34.

[137] ECHR (n 132) art 8.

[138] *Catt v the United Kingdom* App no 43514/15 (ECtHR, 24 January 2019) para 9.

[139] Ibid. para 93.

[140] Ibid.

[141] Ibid.

[142] Ibid. para 123.

[143] Ibid.

lection of information on individuals 'solely on the basis that they belong to particular movements or organisations which are not proscribed by law' should be generally 'prohibited'.[144]

The Court's conclusion in Catt is consistent with its earlier judgment in *Segerstedt-Wiberg and Others v Sweden*, where the Court similarly rejected the Swedish government's argument that because information gathered on 'dissidents' had 'emanated from open sources', and 'public domain' information, there had been no interference with Article 8 ECHR.[145] While in *Shimovolos v Russia*, delivered by a unanimous Court in 2021, the Court held that government collection of data and information even 'exclusively' on a person's 'public activities' is an interference with the right to private life under Article 8 ECHR.[146]

Indeed, the Court has also delivered numerous judgments on the use of automated systems and data collection systems used for government surveillance. Crucially, the ECtHR has emphasised that surveillance systems using 'automated and systemic data collection' had 'reached a level of sophistication which is hardly conceivable for the average citizen'.[147] Indeed, the Court warned about the capacity of governments to acquire 'detailed profile[s] of the most intimate aspects of citizens' lives', which may result in 'particularly invasive' interferences with private life.[148] Similarly, the Court has found a violation of Article 8 over a system allowing storing of a person's photograph in a police database, where the police could apply facial recognition and facial mapping techniques to the image.[149] The Court emphasised the essential importance of Article 8 to guard against the 'risk of arbitrariness' which flows from vesting 'obscure' powers with the State, and 'especially where the technology available is continually becoming more sophisticated'.[150]

IAs such, it is quite clear that this idea of trawling the online environment for harmful content, amassing databases, and sharing this data, raises serious questions under Article 8.[151] Importantly, the Court's emphasis and concern for protecting political expression from the chilling effect is particularly relevant for the monitoring and gathering of social media content.

[144] Ibid. para 124.
[145] *Segerstedt-Wiberg and Others v Sweden* App no 62332/00 (ECtHR, 6 June 2006) para 71.
[146] *Shimovolos v Russia* App no 30194/09 (ECtHR, 21 June 2011) para 65.
[147] *Szabó and Vissy v Hungary* App no 37138/14 (ECtHR, 12 January 2016) para 68.
[148] Ibid. para 70.
[149] *Gaughran v the United Kingdom* App no 45245/15 (ECtHR, 13 February 2020) para 70.
[150] Ibid. para 86.
[151] On the point of data sharing between public authorities, see also case law from the Court of Justice of the European Union under Art 7 (right to respect for private life) and Art 8 (protection of personal data) of the EU Charter, particularly Case C-311/18 *Data Protection Commissioner v Facebook Ireland Limited and Maximillian Schrems*, EU:C:2020:559. See E Kosta and I Kamara, 'The Right to an Effective Remedy in International Data Transfers of Electronic Evidence: Past Lessons and Future Outlook' (2023) 16 *Review of European Administrative Law* 57.

### 4.2. Freedom of expression and government leveraging of platforms

The second question is whether the leveraging of platforms for the monitoring of online expressive activity, and the removal of online, under the DSA is compatible with the right to freedom of expression under both international and European human rights law. As noted above, this can take the form of government-platform initiatives, where platforms remove content following referrals from a government or international agency, or non-judicial authorities ordering the removal of content.

First, under international freedom of expression standards, government measures seeking to leverage platforms for the removal of content are quite suspect. UN human rights bodies, when interpreting Article 19 ICCPR, which guarantees freedom of expression, have been quite clear on the fundamental principle that governments 'should only seek to restrict content pursuant to an order by an independent and impartial judicial authority, and in accordance with due process and standards of legality, necessity and legitimacy'.[152] Further, governments are required to 'refrain' from adopting regulation where 'government agencies', rather than 'judicial authorities', become the 'arbiters of lawful expression', and States should avoid 'delegating responsibility to companies as adjudicators of content'.[153]

Indeed, and as alluded to earlier, there is considerable concern among UN human rights bodies over the path being taken in Europe in this regard. The UN Special Rapporteur on freedom of expression has warned about how government authorities are 'increasingly' seeking to remove online content 'outside of legal process or even through terms of service requests', giving the example of specialist EU units, including the Internet Referral Unit, in order to 'cooperate with platforms to remove' content.[154] The UN Special Rapporteur has also warned how States put 'pressure' on platforms to 'accelerate content removals through non-binding efforts', with the Rapporteur criticising that 'most of which have limited transparency'.[155] There was also criticism of EU governments using trusted flagger mechanisms,[156] with the UN Special Rapporteur criticising how '[a]rrangements to coordinate content actions

[152] *Gaughran v the United Kingdom* (n 153) para 66.

[153] Ibid.

[154] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc. A/HRC/38/35 (6 April 2018), para 68.

[155] Ibid.

[156] The UN Special Rapporteur has criticised the mechanism of trusted flaggers, including 'government flaggers', as '[f]lags typically do not enable nuanced discussions about appropriate boundaries', and there is '[t]here is little or no public information' on the 'interpretations of legal or community standards or their influence over company decisions' (UN Doc A/HRC/38/35, 6 April 2018, para 34). See also K Crawford and T Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint' (2016) 18 *New Media & Society* 410. See also European Commission, 'Code of Conduct on Countering Illegal Hate Speech Online' (30 June 2016), at ec.europa.eu. For a critique, see E Coche, 'Privatised Enforcement and the Right to Freedom of Expression in a World Confronted with Terrorism Propaganda Online' (2018) 7 *Internet Policy Review* 1; and 'EU Fails to Protect Free Speech Online, Again' (Article 19, 28 September 2017), at www.article19.org.

with State input exacerbate concerns that companies perform public functions without the oversight of courts and other accountability mechanisms'.[157] Thus, the proliferation of platform-government initiatives that turn government agencies and law enforcement bodies into arbiters of lawful and unlawful expression are quite questionable under international standards on freedom of expression; and the subsequent delegation of decision about free expression to platforms.

Similar to the position under international human rights law, it must be noted that European human rights law on freedom of expression also places a considerable question mark over these measures. In this regard, Article 10 ECHR guarantees freedom of expression,[158] and importantly, there has been recent case law from the ECtHR, specifically on online platforms and legislation which allow non-judicial authorities to order the removal of content, and imposing short time-frames for removal of content. Further, the Council of Europe's Committee of Ministers has explicitly stated that 'State authorities should obtain an order by a judicial authority or other independent administrative authority, whose decisions are subject to judicial review, when demanding intermediaries to restrict access to content'.[159]

First, one of the most important cases is *Kablis v Russia*,[160] where the European Court considered Russia's Information Act, which requires online platforms to remove certain illegal content, including 'extremist' content, 'within twenty-four hours' by order of an administrative authority. Importantly, the Court unanimously found that a blocking order issued over a post on Russia's largest social media platform VKontakte (VK) because it was 'illegal' content, violated the right to freedom of expression. Crucially, the Court held that the blocking measure was should be classified as a 'prior restraint', as it was 'taken before a judicial decision was issued on the illegality of the published content'.[161] Prior restraints are the most suspect restrictions on free expression under Article 10 ECHR, and as such, the Court applies its highest standard of scrutiny – 'most careful scrutiny' – to prior restraints, and consequently, found the order violated Article 10 ECHR. Notably, this was so, even where under the legislation there existed judicial review, with the Court unanimously finding that the legislation lacked the 'necessary guarantees against abuse' and to protect against the 'dangers inherent in prior restraints'.[162] Thus, legislation that allows non-judicial authorities to order platforms to remove content within tight-time

[157] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc. A/HRC/38/35 (6 April 2018), 20.

[158] Art 10 of the ECHR.

[159] Recommendation CM/Rec(2018)2 (n 94) section 1.3.2. See also Recommendation CM/Rec(2022)13 of the Committee of Ministers to member States (6 April 2022) on the impacts of digital technologies on freedom of expression, section 1.1.

[160] *Kablis v Russia* App nos 48310/16 and 59663/17 (ECtHR, 30 April 2019).

[161] Ibid. para 90.

[162] Ibid.

frames are classed as prior restraints by the European Court, and subject to the Court's strict test under Article 10 ECHR.

Second, not only does the European Court apply its strictest standards of scrutiny to fast removal orders by non-judicial administrative authorities, the Court has also found violation of Article 10 where courts make sure orders. Notably, the Court has delivered a number of judgments on a provision in Poland's election legislation, which allows a court to order prohibit or restrain publication of 'untrue data or information', with the court required to examine the application 'within twenty-four hours'.[163] In its most recent *Brzeziński v Poland* judgment in 2019, a unanimous European Court found that an order under the provision violated Article 10, which had been issued against a local politician during an election.[164] Crucially, the Court held that the procedure 'effectively deprived' the politician of the 'protection afforded by Article 10', where the domestic courts had near-immediately classified the information as 'untrue'.[165] Indeed, it two other judgments on the same provision, the Court also found violation of Article 10. The Court emphasised the 'very short time-limits', and in particularly trenchant language, held that the 'fairness of the proceedings may be called into question'[166] While in *Kita v Poland*, the Court also unanimously held there had been a violation of Article 10 over the 'untrue information' proceedings, finding the national courts 'unreservedly qualified all of [the statements] as statements which lacked any factual basis', and the 'standards applied' by the national courts were 'not compatible with the principles embodied in Article 10'.[167]

### 4.3. Targeting expression based on broad categorisation of content as illegal

The third issue that must be addressed is that a great deal of government measures leveraging platforms to restrict expression are premised on the assumption that when a piece of content simply falls within a prohibited category of content, it can be removed. And similarly, government surveillance of platforms to root illegal and harmful content is simply a matter of identity content that falls within a certain category of content. However, it is never the case under human rights law that just because shocking, disturbing or offensive expression is labelled 'terrorist propaganda', 'hate speech', 'disinformation', or 'extremist material', by law enforcement, public officials, or public prosecutors, that such content is unquestionably illegal, and it is perfectly fine to restrict a person's free expression. It is always important to bear in mind that human rights courts and bodies, such as the ECtHR, regularly find that

[163] *Brzeziński v Poland* App no 47542/07 (ECtHR, 25 July 2019) para 28.
[164] Ibid.
[165] Ibid. para 58.
[166] *Kwiecień v Poland* App no 51744/99 (ECtHR, 9 January 2007) para 55.
[167] *Kita v Poland* App no 57659/00 (ECtHR, 8 July 2008) para 51.

national prosecutions and convictions for 'terrorist propaganda', 'hate speech', and 'false information', violate the right to freedom of expression.[168]

Indeed, when the ECtHR recently found that an activist's prosecution for incitement to hatred violated the right to freedom of expression, the Court took the opportunity to stress that it is 'vitally important' that European governments adopt a 'cautious approach', and 'strictly construe' incitement laws in order to avoid 'excessive' interference with free expression 'under the guise of action taken against "hate speech"'.[169] Importantly, in the area of alleged 'terrorist' content, the Court has similarly displayed great suspicion of near-automatic convictions imposed by national courts for 'terrorist propaganda' over publication of statements from proscribed terrorist organisations, and where national courts effectively ignore whether the actual content meets the strict criteria for incitement to violence under freedom of expression law.[170] Crucially, in its unanimous 2018 judgment in *Savva Terentyev v Russia*,[171] the Court reiterated its strict test under Article 10 for incitement to violence: courts must be satisfied that *a)* there is a 'clear and imminent danger' of violence resulting from the expression; *b)* the 'intention' of the speaker was to incite violence; and *c)* careful consideration the 'context' of the expression.[172] As such, even expression with a 'violent tone' can be protected political expression, such as statements that 'political power grows out of the barrel of the gun' and 'it is the barrel of the gun that will call into account', as Article 10 ECHR has long protected shocking, offensive and disturbing expression.[173]

The foregoing reflects a fundamental and overacting principle of freedom of expression law under Article 10 ECHR: it is never simply the case that because a piece of content falls within a specific legal provision on illegal content that it automatically follows such content can be legitimately restricted consistent with Article 10's guarantee of free expression. Far from it. The Court has long held that it is '*not* sufficient' that an interference with free expression is imposed 'because its subject matter fell within a particular category or was caught by a legal rule formulated in general or absolute terms'.[174] What is strictly 'required' under Article 10 ECHR is a 'detailed judicial assessment' that the specific interference with free expression was 'necessary in the specific circumstances',[175] and applying the strict standard of review under Article 10 ECHR. This is not a fringe principle, and recently reaffirmed

[168] See for example *Özer v Turkey (No 3)* App no 69270/12 (ECtHR, 11 June 2020); and *Savva Terentyev v Russia* App no 10692/09 (ECtHR, 28 August 2018).

[169] *Stomakhin v Russia* App no 52273/07 (ECtHR, 9 May 2018) para 117.

[170] *Gözel and Özer v Turkey* App nos 43453/04 and 31098/05 (ECtHR, 6 July 2010).

[171] *Savva Terentyev v Russia* (n 172).

[172] Ibid. paras 70, 73, and 84.

[173] *Gül and Others v Turkey* App no 4870/02 (ECtHR, 8 June 2010) para 41.

[174] *Kablis v Russia* (n 164) para 103, applying *The Sunday Times v the United Kingdom* App no 6538/74 (ECtHR, 26 April 1979) para 65, emphasis added.

[175] *Perinçek v Switzerland* App no 27510/08 (ECtHR, 15 October 2015) para 275.

by the 17-judge Grand Chamber of the European Court in its seminal *Perinçek v Switzerland*.[176]

Indeed, in the *Kablis* judgment, the Court also held that just because content 'breached a statutory prohibition',[177] it does not follow that the content can simply be blocked. Instead, it must be determined under Article 10 ECHR whether it was 'necessary in a democratic society to block the publications at issue, having regard to the facts and circumstances of the case'.[178] The European Court castigated the national courts for failing to engage in this strict standard of review, and the Court held that there had been a violation the right to freedom of expression as the post 'did not create any real risk of public disorder' or 'harmful consequences for public safety'.[179]

There is a danger that the type of government measures currently being implemented and rolled out to tackle supposed illegal and harmful content fundamentally offend the principles of Article 10 ECHR mentioned above. These government measures, whether widespread government monitoring of allegedly illegal and unlawful content, government flagging of such content, and government-platforms cooperation on removal of content, are all based on the assumption that simply identifying a piece of content as allegedly falling within a prohibition contained in a legal provision means that all bets are off, and such content can be requested to be blocked, outright blocked, or the content gathered and stored in databases of harmful content.

## 5. Conclusion

A number of concluding points can be made about the current proliferation of government leveraging of platforms for the purposes of surveillance and removal of content, and the operation of the DSA more generally. First, there are serious concerns under international and human rights law on European and national authorities engaging in monitoring and data-gathering of user content on online platforms. Second, law enforcement involvement in online monitoring and removal of content is at odds with the principle under human rights law that should only seek to restrict content pursuant to an order by a court. Third, there are further serious question marks over non-judicial authorities ordering content to be removed. Fourth, it must be emphasised that European governments need to take the danger of the chilling effect seriously. All these measures are incentivising platforms to removing content rather than risk liability or punishment. And most important, widespread online surveillance is

[176] Ibid. para 275. See also *The Sunday Times v the United Kingdom* (n 178) para 65 (holding that 'neither is it sufficient that the interference was imposed because its subject-matter fell within a particular category or was caught by a legal rule formulated in general or absolute terms: the Court has to be satisfied that the interference was necessary having regard to the facts and circumstances prevailing in the specific case before it').

[177] *Kablis v Russia* (n 164) para 103.

[178] Ibid. para 103.

[179] Ibid. para 105.

creating an online environment where individuals will refrain from posting important content for fear of government surveillance.

The purpose of this article has not been to argue that the DSA is, per se, a violation of the right to freedom of expression or the right to privacy. However, its central thesis, and contribution to the literature, is that the operationalisation of the DSA by various actors, including the European Commission, national courts, and national Digital Services Coordinators, must be consistent with the rights to freedom of expression and privacy under both international and ECHR case law, in particular the principle of proportionality. This insight further invites literatures on content moderation, digital constitutionalism and the governance of information dissemination to deep dive into the actual implementation of legislation and the field of possibility – especially those for enhance online surveillance – that they open, and this even when legislation aims at ensuring the governance of those actors, like platforms, that currently look like particularly powerful ones.

Finally, a recent worrying controversy involving the European Commission completely reinforces this article's argument about the need to pay close attention to the implementation of such a far-reaching and comprehensive legislation and its consistency with fundamental rights. In the context of protests and rioting in France during summer 2023, and the French president raising the possibility of 'blocking social media platforms during civil unrest', the European Commissioner for the Internal Market, Thierry Breton, made what were described as 'incendiary statements', suggesting online platforms could indeed be 'blocked' under the DSA for failing to remove 'hateful content'.[180] Following a particularly strong backlash from free-expression groups, including an open letter signed by 60 civil society groups decrying the Commissioner's comments, the Commissioner was forced to issue a detailed clarification. The Commissioner explicitly admitted that any measures taken under the DSA needing to be 'proportionate', and could only be taken 'in line with international law and jurisprudence by the European Court of Human Rights'.[181] This episode unfortunately is the danger associated with the DSA, and this article seeks centres the right to freedom of exposit and privacy in its implementation.

[180] Access Now, *DSA is Not a Censorship Tool: Commissioner Breton must Clarify Blocking Statement* (2 August 2023), at www.accessnow.org.

[181] See Statement by the European Commissioner for Internal Market, Thierry Breton published at Access Now, *Civil Society Gets its Confirmation from EU Commissioner: No Internet Shutdowns under DSA* (2 August 2023), at www.accessnow.org.