cerre Centre on Regulation in Europe

SYSTEMIC RISK IN DIGITAL SERVICES: BENCHMARKS FOR EVALUATING MANAGEMENT OF RISK OF TERRORIST CONTENT DISSEMINATION

ISSUE PAPER

November 2024

Sally Broughton Micova

Issue Paper

Systemic Risk in Digital Services: Benchmarks for Evaluating Management of Risk of Terrorist Content Dissemination

Sally Broughton Micova

November 2024

As provided for in CERRE's bylaws and procedural rules from its "Transparency & Independence Policy", all CERRE research projects and reports are completed in accordance with the strictest academic independence.

The project, within the framework of which this report has been prepared, received the support and/or input of CERRE member organisations. However, they bear no responsibility for the contents of this report. The views expressed in this CERRE report are attributable only to the authors in a personal capacity and not to any institution with which they are associated. In addition, they do not necessarily correspond either to those of CERRE, or of any sponsor or of members of CERRE.

info@cerre.eu – www.cerre.eu

Table of Contents

About CERRE

Providing high quality studies and dissemination activities, the Centre on Regulation in Europe (CERRE) is a not-for-profit think tank. It promotes robust and consistent regulation in Europe's network, digital industry, and service sectors. CERRE's members are regulatory authorities and companies operating in these sectors, as well as universities.

CERRE's added value is based on:

- its original, multidisciplinary and cross-sector approach covering a variety of markets, e.g., energy, mobility, sustainability, tech, media, telecom, etc.;
- the widely acknowledged academic credentials and policy experience of its research team and associated staff members;
- its scientific independence and impartiality; and,
- the direct relevance and timeliness of its contributions to the policy and regulatory development process impacting network industry players and the markets for their goods and services.

CERRE's activities include contributions to the development of norms, standards, and policy recommendations related to the regulation of service providers, to the specification of market rules and to improvements in the management of infrastructure in a changing political, economic, technological, and social environment. CERRE's work also aims to clarify the respective roles of market operators, governments, and regulatory authorities, as well as contribute to the enhancement of those organisations' expertise in addressing regulatory issues of relevance to their activities.

About the Authors

Sally Broughton Micova is a CERRE Academic Co-Director and an Associate Professor in Communications Policy and Politics at the University of East Anglia (UEA). She is also a member of UEA's Centre for Competition Policy.

Her research focuses on media and communications policy in Europe.

She completed her PhD in the Department of Media and Communications at the London School of Economics and Political Science (LSE), after which she was an LSE Teaching and Research Fellow in Media Governance and Policy and Deputy Director of the LSE Media Policy Project.

Acknowledgements

The authors are extremely grateful for the substantive input from fellow project team members Daniel Schnurr from the University of Regensburg and Andrea Calef and Bryn Enstone, both from the UEA Centre for Competition Policy (CCP). She also thanks Peter Courridge and Juliette Hardman of CCP for their assistance with the executing the empirical investigation and all those who gave their time to participate as interviewees, as well as the steering committee members and peer reviewer for their comments.

1. Introduction

The European Union's Digital Services Act (DSA) requires all intermediary services in scope to act against illegal content when the receive and order from a judicial or administrative authority. This requirement can be seen as simply giving more detail and arguably strength to the over two-decade old expectation in the e-Commerce Directive that intermediaries action illegal content brought to their attention. A significant innovation in the DSA is that designated very large online platforms and search engines (VLOPs and VLOSEs) are also required to assess and mitigate the systemic risk of the dissemination of illegal content through their services. As has been argued in previous CERRE reports, this implies looking into relationships and interactions within wider ecosystems and not simply assessing the speed and accuracy of removals in response to reports.[1] With regulators, researchers, civil society and the general public finally due to see the public versions of the first attempts by VLOP and VLOSE providers at systemic risk assessments, this issue paper aims to contribute to the evaluation of the DSA's systemic risk management approach to the systemic risk of the dissemination of illegal content.

Since the DSA defines illegal content broadly, this issue paper focuses on the specific case of terrorist content in order to make two contributions. Firstly, it proposes benchmarks for assessing the mitigation of risk that are specific to terrorist content. Secondly, it suggests lessons that might be drawn from efforts to prevent the dissemination of terrorist content that are relevant for the wider challenge of illegal content.

Terrorist content is of particular interest because it is also covered by harm specific EU legislation, the 2021 Regulation on addressing the dissemination of terrorist content online (TERREG), and several transnational initiatives aimed at combatting it. The potential harms from the dissemination of terrorist content identified in the TERREG fall into two categories. The first includes the potential for radicalisation of individuals or incitement leading to the commission of terrorist acts, threats to public security and even loss of life. The second category of potential harm acknowledged in the TERREG is impingement of fundamental rights, especially rights to expression and information, from measures aimed at combatting terrorist content.[2] The DSA's emphasis on proportionality and multiple references to fundamental rights in relation to the treatment of illegal content also indicates this dual pronged understanding of the harm to be prevented.[3]

The proposals presented in this issue paper are based on examination of the DSA, the TERREG, other EU documents on the issue of terrorist content, documents from the UN, the OSCE and G20 on the issue, reports and position papers from various initiatives and civil society organisations, and academic literature as well as on expert interviews with practitioners engaged in efforts to combat terrorist

[1] Sally Broughton Micova and Andrea Calef, 'Elements for Effective Systemic Risk Assessment under the DSA', CERRE Research Reports (Brussels: Centre on Regulation in Europe (CERRE), 17 July 2023), https://cerre.eu/publications/elements-for-effective-systemic-risk-assessment-under-the-dsa/; Sally Broughton Micova et al., *Cross-Cutting Issues for DSA Systemic Risk Management: An Agenda for Cooperation* (Brussels: Centre on Regulation in Europe asbl (CERRE), 2024), https://cerre.eu/publications/cross-cutting-issues-for-dsa-systemic-risk-management-an-agenda-for-cooperation/.

[2] Regulations 2021/784 on addressing the dissemination of terrorist content online (TERREG) Recital 10.

[3] E.g. Regulation 2022/2065 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (DSA) Recitals 22 & 26.

content and civil society representatives. The paper first elaborates some of the key definitions and definitional issues, and then describes the actors and vulnerabilities within integrated ecosystem of these actors. It then suggests four benchmarks for evaluating management of the systemic risk of terrorist content dissemination in two areas, internal mitigation and wider collaboration. It proposes benchmarks on content exposure and on fundamental rights protection for the measures taken to mitigate terrorist content dissemination on their services. It also proposes two benchmarks for evaluating each service's engagement in collaboration and contribution to mitigation in the wider ecosystem. The paper then discusses some metrics and measures that could be used to assess whether those benchmarks are being met, identifying gaps in the existing reporting data. It calls for more nuanced standardisation, greater attention to accuracy measures, and increased transparency on the roles of various actors.

The issue paper concludes by arguing that there is a need:

- to balance removal or exposure prevention targets with fundamental rights targets in relation to illegal content;
- for an inclusive and coherent approach to evaluating collaboration and knowledge sharing efforts;
- greater transparency and assessment of the roles and relationships with law enforcement and other Member State authorities.

With the public versions of the first systemic risk assessments and audit reports expected to be released before the end of 2024, the paper suggests **five specific issues that should be examined across the systemic risk reports to enable evaluation against the benchmarks proposed for combatting terrorist content**. It argues that evidence from the VLOPs and VLOSEs will need to be combined with evidence from reporting by smaller services to thoroughly evaluate the mitigation of systemic risk of terrorist content dissemination.

2. Definitional Issues

The DSA's broad definition of illegal content presents a challenge because it introduces a need to determine when Member State law might conflict with Union law.

Article 3(h) states: "'illegal content' means 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;"

Any information that relates to an activity, product or service that is illegal in a Member State may be in scope as long as the Member State law does not conflict with EU law. Definitions for illegal hate speech, for example, are not fully harmonised despite the 2008 Framework Decision on Racism and Xenophobia.[4] One can imagine scenarios in which conflicts between Union and Member state law might occur, however, there is a comparatively high level of harmonization in the definition of terrorist content.

The EU has harmonised the definition of terrorism across Member States and established a legal definition of terrorist content through the TERREG. Establishing legal certainty with a definition for terrorist content was one of the aims of the Regulation and a particularly contentious issue during the protracted negotiations over the TERREG.[5] The result of these negotiations is relatively precise wording in Article 2.7:

> *'terrorist content' means one or more of the following types of material, namely material that:*
>
> *(a) incites the commission of one of the offences referred to in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541, where such material, directly or indirectly, such as by the glorification of terrorist acts, advocates the commission of terrorist offences, thereby causing a danger that one or more such offences may be committed;*
>
> *(b) solicits a person or a group of persons to commit or contribute to the commission of one of the offences referred to in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541;*
>
> *(c) solicits a person or a group of persons to participate in the activities of a terrorist group, within the meaning of point (b) of Article 4 of Directive (EU) 2017/541;*
>
> *(d) provides instruction on the making or use of explosives, firearms or other weapons or noxious or hazardous substances, or on other specific methods or techniques for the purpose*

[4] The 2008 Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178 requires the criminalisation of public incitement to violence or hatred based on race, colour, religion, descent or national or ethnic origin in Member State law, but hate crimes and hate speech remain subjects of Member State law rather than EU Law, which has partly inspired recent calls from MEPs and the Commission for them to be adopted by the Council into the list of "EU crimes." (see https://www.europarl.europa.eu/legislative-train/theme-a-new-push-for-european-democracy/file-hate-crimes-and-hate-speech).

[5] Reem Ahmed, 'Negotiating Fundamental Rights: Civil Society and the EU Regulation on Addressing the Dissemination of Terrorist Content Online', *Studies in Conflict & Terrorism*, n.d., 1–25, https://doi.org/10.1080/1057610X.2023.2222890.

of committing or contributing to the commission of one of the terrorist offences referred to in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541;

(e) constitutes a threat to commit one of the offences referred to in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541;

The TERREG's definition of terrorist content clearly relies on the definitions of terrorist groups and terrorist offences in the EU's 2017 Directive on combatting terrorism.[6] Here is where some vagaries and definitional challenges arise. The Directive provides a list of dangerous or violent acts that are considered terrorism if done with the aim of "seriously intimidating a population; (b) unduly compelling a government or an international organisation to perform or abstain from performing any act; or (c) seriously destabilising or destroying the fundamental political, constitutional, economic or social structures of a country or an international organisation."[7] Under EU law, therefore, terrorist content is that which incites or solicits people to those purposes, provides instructions for the tools to commit acts for those purposes, or solicits them to join a terrorist group formed to commit acts for those purposes.[8]

The definition of terrorist offences has been criticised for being problematic because it includes acts that might be conducted for purposes other than terrorism, and even purposes that could be legitimate in some cases, such as compelling a government to abstain from an act.[9] It is not limited to acts that are intentionally violent to members of the population as was recommended by the UN Special Rapporteur on the issue,[10] but instead includes acts on infrastructure and property. As Rojszczak has argued, the need to identify a purpose to the dissemination of the content stemming from the 2017 Directive, also introduces ambiguity and potential for over-identification of terrorist content.[11] These lines regularly need to be drawn in relation to climate and environmentalist protest. Greenpeace, for example, is known for seizure of vessels and disrupting the supply of what might be considered a fundamental natural resource, both acts listed in the Directive's definition. Greenpeace activists so far have faced trespassing, public disorder and even piracy charges in various countries, as well as civil suits.[12] However, as Human Rights Watch has documented, environmental activists have

[6] Directive (EU) 2017/541

[7] Directive (EU) 2017/541 Article 3.2

[8] Directive (EU) 2017/541 Article 2 defines terrorist groups as "a structured group of more than two persons, established for a period of time and acting in concert to commit terrorist offences"

[9] Tarik Gherbaoui and Martin Scheinin, 'A Dual Challenge to Human Rights Law: Online Terrorist Content and Governmental Orders to Remove It', *Journal Européen Des Droits de l'homme-European Journal of Human Rights* 1 (2023): 3–29.

[10] Martin Scheinin, 'Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism, Martin Scheinin: Ten Areas of Best Practices in Countering Terrorism' (New York: UN Human Rights Council, 22 December 2010), https://www2.ohchr.org/english/bodies/hrcouncil/docs/16session/a-hrc-16-51.pdf.

[11] Marcin Rojszczak, 'Gone in 60 Minutes: Distribution of Terrorist Content and Free Speech in the European Union', *Democracy and Security* 20, no. 2 (2 April 2024): 179–209, https://doi.org/10.1080/17419166.2023.2250731.

[12] https://www.ft.com/content/3041e388-f1a5-4bee-a756-fa58e8fd639c; https://www.ntu.ac.uk/about-us/news/news-articles/2023/07/expert-blog-obviously,-they-are-not-pirates-the-european-court-on-human-rights-rules-in-favour-of-greenpeace-activists-in-the-arctic-sunrise-case

been classed as terrorists or threats to security under counter-terrorism measures in multiple countries, including some EU Member States.[13]

The boundaries of what constitutes a terrorist offence will be determined by Member State courts, or the CJEU, because the definitions of terrorist offences and terrorist groups in the 2017 Directive were designed to be transposed into Member State criminal law. When people are accused of terrorist offences, a court decides whether their act falls into that category, weighing up freedom of expression and other rights. However, the definition of terrorist content is in an EU Regulation that empowers Member State authorities with an administrative mechanism, one that gives the designated *competent authorities* in Member States the power to require content to be removed within an hour, without any trial or equivalent.[14] These competent authorities decide what content qualifies for a removal order and digital services decide in the context of the specific measures they take to prevent dissemination if the services is considered exposed to terrorist content.[15] There are redress mechanisms proscribed in both the TERREG and the DSA but these are non-public processes for appeal individual content decisions. This raises questions about who is drawing the lines between illegal and what the industry refers to as *borderline content,* as well as what kind of oversight there is on how those are being drawn.

The term borderline content is an ambiguous umbrella term for a vast array of content that is not illegal and may not breach services' terms yet is a cause for concern. Such content is often de-amplified by recommender systems or otherwise actioned with implications for expression rights.[16] In the context of terrorism, this might include the kind of content used as *beacons* to entice users into communities on messaging apps or other less moderated platforms, coded messages, manifestos, or content aimed at radicalising that falls short of the clear solicitation or incitement required in the TERREG's definition of terrorist content. In practice, the terms and conditions of the online platforms, especially those designated as VLOPs, prohibit much broader categories of content than that defined by TERREG. This can include *adjacent content* that overlaps with other harmful categories such as hate speech, violent & graphic content, harassment and even some disinformation.[17] Recent work within Global Internet Forum to Counter Terrorism (GIFCT) and the EU Internet Forum has attempted to clarify the types and borderline content and arrive at some common understandings of this necessarily dynamic category of content.[18]

The OECD's fourth investigation into transparency reporting on terrorist and violent extremist content online found that there has been progressively more detail in the categories banned by the 50 most

[13] Letta Tayler and Cara Schulte, 'Targeting Environmental Activists With Counterterrorism Measures Is an Abuse of the Law', *Human Rights Watch* (blog), 29 November 2019, https://www.hrw.org/news/2019/11/29/targeting-environmental-activists-counterterrorism-measures-abuse-law.

[14] Gherbaoui and Scheinin, supra note 9.

[15] These are set out in Art 5 TERREG, supra note 2.

[16] Stuart Macdonald and Katy Vaughan, 'Moderating Borderline Content While Respecting Fundamental Values', *Policy & Internet* 16, no. 2 (1 June 2024): 347–61, https://doi.org/10.1002/poi3.376.

[17] As explained by interviewee CS Part 4, there are overlaps in the possible reasons services action content. Terrorist content, or borderline terrorist content can often be actioned by services because of violations in these adjacent categories.

[18] GIFCT, *Borderline Content: Understanding the Gray Zone* (2023)

popular platforms, noting several examples from those also designated as VLOPs.[19] However, they also noted two VLOPs that had reduced clarity and removed clearly defined categories of terrorism or terrorist content from their policies. Therefore, the points at which content is permissible can vary considerably across services and are often much more inclusive than the stricter EU definition of terrorist content. Importantly, they are set by the service providers themselves.

Journalistic content has long posed a challenge to establishing the borders of terrorist content because terrorism itself is, as Borelli notes, "a propaganda of the weak" that relies heavily on media.[20] Terrorist acts are committed to attract attention, especially media attention, and reach large audiences. This makes achieving the appropriate balance to prevent both harm from the effects of terrorism and harm from excessive restrictions on expression rights difficult. For example, a Tech Against Terrorism investigation content from the livestream of the attacker who killed 10 people in Buffalo NY in 2022 remained highly discoverable on major platforms, despite it being stopped within two minutes by the original source. They attributed this, in part, to the large amount of content from the livestream that had been edited and branded by local and regional media outlets, complicating automatic detection.[21]

The TERREG explicitly excludes content that is disseminated for "educational, journalistic, artistic or research purposes" from being considered terrorist content, and states that "an assessment shall determine the purposes of the dissemination."[22] This introduces another requirement for judgements to be made about the purposes of the dissemination of the content, either by competent authorities issuing removal orders or by services implementing automatic detection and removal mechanisms.

There is convincing evidence that digital services rely heavily on designation lists from the UN Consolidated list, or national governments, as a proxy for purpose.[23] It has been noted that the designation lists are often skewed towards Islamic terrorism, insufficiently cover far-right groups, and are inadequate for lone actor cases.[24] They can help with determining the boundaries between illegal and borderline content and in ensuring journalistic content is excluded, however they can also introduce biases.

The definition of terrorist content may be clearer and more harmonised than some other types of illegal content. Nevertheless, there are vulnerabilities stemming from the fact that, as is normal in criminal law, a determination of purpose or intent is required. Instead of being done by a court based on evidence gathered by law enforcement, this must be made by the VLOPs and VLOSEs, very quicky

[19] OECD, 'Transparency Reporting on Terrorist and Violent Extremist Content Online: Fourth Edition', OECD Digital Economy Papers (Paris: OECD Publishing, 2024), https://doi.org/10.1787/901cb8cf-en.
[20] Marguerite Borelli, 'Social Media Corporations as Actors of Counter-Terrorism', *New Media & Society* 25, no. 11 (2023): 2877–97.
[21] Tech Against Terrorism, 'State of Play 2022' (Tech Against Terrorism, 2022), https://www.techagainstterrorism.org/hubfs/FINAL-State-of-Play-2022-TAT.pdf.
[22] TERREG Art 1.3
[23] Tech Against Terrorism, 'Who Designates Terrorism' (London: Tech Against Terrorism, March 2023), https://techagainstterrorism.org/hubfs/TAT-Designation-Report-March-2023.pdf; Chris Meserole and Daniel Byman, 'Terrorist Definitions and Designations Lists: What Technology Companies Need to Know', Global Research Network on Terrorism and Technology (London: Royal United Services Institute for Defence and Security Studies (RUSI), 2019), https://www.brookings.edu/wp-content/uploads/2019/07/GRNTT-Paper-No.-7.pdf.
[24] GIFCT (2024) Risk Mitigation in Applying Government Terrorist Designation Lists https://def-frameworks.gifct.org/risk-mitigation-supplement/

and often using automated means. This introduces risk to fundamental rights. Reliance on designation lists as a proxy for that determination of purpose has limits and brings risks of under-identification, bias, and political manipulation. The DSA's Recital 84 instructs that "when assessing the systemic risks identified in this Regulation, those providers should also focus on the information which is not illegal but contributes to the systemic risks identified in this Regulation." This means that, unlike the TERREG which concerns only the illegal content itself, in the DSA there is an expectation that the assessment and mitigation of risk from terrorist content includes addressing borderline content that may facilitate or encourage its dissemination.

Exactly what qualifies as borderline terrorist content and what identifies tactics such as beaconing and coding will vary across different types of platforms. However, the evidence from this investigation indicates that some common understandings could be reached on distinctions to be made, groups affected by actioning of content, and exclusion criteria, even though these would need to be regularly revisited. Without a mechanism for inclusive discussions on dynamic understandings of borderline content, that at least involve the array of prevention-focused actors and representatives of groups affected by mitigation measures, there are risks both of failing to prevent dissemination and of over-reach at the expense of freedom of expression and other rights.

3. Integrated ecosystem around terrorist content

It is evident from the accounts of those interviewed and the reports of various organisations engaged in countering online terrorist content that VLOPs and VLOSEs are only part of a complex ecosystem of actors involved in the dissemination of such content and efforts to counter it. As argued in a previous CERRE report, the assessment of systemic risk necessarily involves reflection on relationships with other actors and on the specific contribution of a service to the risk.[25] This section, therefore, sets out the various other actors involved and discusses the vulnerabilities arising within the system.

3.1 The interlinked actors

Designated very large services – A very diverse set of services have been designated under the DSA as VLOPs and VLOSEs. All could be susceptible to some kind of terrorist activity, but they are not equally exposed to risks of the dissemination of terrorist content per se. VLOPs that are online retail services or app stores are not primarily engaged in the dissemination of content, while social media and video-sharing platforms have that as their core functionality.[26] VLOSEs are not used to disseminate content by users, but are crucial to the dissemination of content as they enable users to find content. Terrorists and sympathisers use different services for different purposes just like any other user. Some are used to reach wide audiences and others for more private communication among sympathisers. As one interviewee pointed out, "also a lot of other online operations exploited by bad actors go above and beyond this limited content lens, more towards operationalization network management and procurement of goods."[27] The risks associated with each service and their role in the ecosystem will vary significantly depending on why each is attractive to terrorists and their sympathisers and how they are being used.

Terrorists – Though this category of malicious actors may seem obvious, it has become important in recent years to break down this category because there have been significant changes in the kinds of perpetrators of terrorism and their behaviour online in recent decades. Europol now breaks down terrorists into Jihadist, right-wing, left-wing and anarchists, ethno-nationalist and separatist, and others.[28] The latest report from the Terrorist Content Analytics Platform (TCAP) of Tech against Terrorism divides data into two categories, Islamist and far-right terrorism.[29] A key distinction made is in the extent and nature of the organisations behind the terrorism. Islamist or Jihadist groups and

[25] Sally Broughton Micova and Andrea Calef, 'Elements for Effective Systemic Risk Assessment under the DSA', CERRE Research Reports (Brussels: Centre on Regulation in Europe (CERRE), 17 July 2023), https://cerre.eu/publications/elements-for-effective-systemic-risk-assessment-under-the-dsa

[26] Online retail services that allow comments or reviews on products or enable the sale of print, audio or audiovisual content are exposed to the risk of dissemination. For a breakdown of the types of designated services and risk related concerns see

[27] Interview CS Part 4.

[28] EUROPOL, 'European Union Terrorism Situation and Trend Report 2023 (TE-SAT)' (The Hague, Netherlands: EUROPOL, 19 December 2023), https://www.europol.europa.eu/publication-events/main-reports/european-union-terrorism-situation-and-trend-report-2023-te-sat.

[29] Tech Against Terrorism, 'Transparency Report: Terrorist Content Analytics Platform' (London: Tech Against Terrorism, 22 August 2023), https://techagainstterrorism.org/hubfs/Tech-Against-Terrorism-TCAP-Transparency-Report-2021-2022.pdf.

ethno-nationalists or separatists are usually structured organisations with public presences. They are therefore engaged systematically in recruitment, training and mobilization. Far-right terrorist are often lone actors, radicalised through exposure to multiple sources of narratives, especially online.[30] The different types of terrorist actors make use of digital services in distinct ways and produce terrorist content for varying purposes.

Smaller digital services – There is ample and convincing evidence of the significant role that smaller services and individual websites play in the dissemination of terrorist content online. These can be small, locally used social media or video-sharing platforms, file sharing services, messaging services, even gaming platforms and simple websites. Among those identified as being "intensively" used by terrorists for disseminating content by the OECD are ones for sharing music, for sharing files, platforms operated by individuals, and several platforms created by known terrorist groups or their sympathisers.[31] Some of these will be within the scope of the DSA and TERREG and therefore subject to the notice and takedown and transparency obligations. Others, however, are out of scope or not interested in compliance for ideological reasons or are deliberately supporting terrorist groups.

National Competent Authorities (NCAs) – These are designated by Member States as required by TERREG and have the power to issue removal notices to digital services, intermediaries. Some of these are Member State security services, police forces, or interior/home affairs ministries. Less often have a court or public prosecutors been designated. Many member states had internet referral units (IFUs) set up prior to the adoption of TERREG, which flagged content to services. This system placed the burden on the services to make the final decision on the content. Where IFUs have been designated as NCA issuing removal orders as well, the pressure to remove simply flagged content may be greater and there is little transparency over this. As Gherbaoui and Scheinin have pointed out, there is a risk of political abuse of removal orders or more subtle forms of overreach due to the power given to the NCAs and their potential lack of expertise in freedom of expression and human rights law.[32]

Europol – The EU's Agency for law enforcement cooperation hosts the European Counter Terrorism Centre, which operates the EU Internet Referral Unit. One of the core tasks of this unit is to flag terrorist content to services for removal, and the TERREG recitals recommend that Member State NCAs use this existing mechanism for notifying services and coordination. This centralises the flagging and removal orders coming from IFUs and other designated NCAs.

The Global Internet Forum to Counter Terrorism – GIFCT was founded by tech companies following the terrorist incident in Christchurch New Zealand and in parallel with the Christchurch Call. As of October 2024, thirty-two of the largest global digital service providers were members, including many, but not all, of the providers of designated VLOPs and VLOSEs. A core element of the cooperation is the Hash-Sharing Database to which members contribute identified content and which they can use to identify and remove content on their services. National governments are also able to add content to this database. GIFCT serves as a forum for knowledge exchange and cooperation around tools, producing insight and resources such as its mapping of the definitions of terrorism and violent

[30] William Allchorn and Katherine Kondor, 'Policing of the Far-Right Online: The Cases of the UK and Hungary', *STUDIES IN CONFLICT & TERRORISM*, 23 March 2023, https://doi.org/10.1080/1057610X.2023.2195063.
[31] OECD, 'Transparency Reporting on Terrorist and Violent Extremist Content Online: Fourth Edition'.
[32] Gherbaoui and Scheinin, 'A Dual Challenge to Human Rights Law: Online Terrorist Content and Governmental Orders to Remove It'.

extremism.[33] It also operates a Content Incident Protocol that is activates when a terrorist incident is being livestreamed.

Tech Against Terrorism – TAT is a non-profit organisation that was founded by the United Nations in 2016. It operates a Terrorist Content Analytics Platform (TCAP) that identifies terrorist content, notifies service providers, and then checks their services to verify effective removal. It currently receives EU funding to make this available for smaller digital services. They are transparent about their inclusion policy, which relies heavily on the designation and sanctions lists of the UN, the EU, the US, UK, Australia and Canada,[34] though is reviewed regularly. They work closely with GIFCT on a mentorship programme to support smaller platforms in meeting the criteria for joining GIFCT.

Commercial insight and content moderation resource suppliers – Several companies have emerged to help digital service providers with efforts to combat terrorist content and compliance with relevant regulations in the EU and other jurisdictions. Some provide resources to help with identifying terrorist content such as databases or lists of groups, use patterns and content features. Some will act as "red teams" to test out services' own identification and removal mechanisms, recommender systems, and other functionalities. A small number of companies came up in the interview data and reports, TRAC operated by Beacham Publishing Corporation, the US-based SITE Intelligence Group Enterprise, academic-run Jihadology, and Moonshot, which was initially founded to combat violent extremism.

Digital Services Coordinators – Although the Commission is responsible for regulating VLOPs and VLOSEs in relation to the management of risks set out in the DSA's Article 34, the Digital Services Coordinators are responsible for enforcing the DSAs general obligations on services addressing illegal content, redress, and reporting. They also must report annually on the number and subject of orders to act against illegal content from their Member State, including terrorist content, and the response to those orders. Therefore, they retain a significant role in relation to preventing the dissemination of terrorist content and generating metrics and other data. Through the Digital Services Board, they also input into the Commission's decisions on designation of services, activation of the crisis mechanism, and regular reporting.

European Commission – The Commission is the regulator for VLOPs and VLOSEs and responsible for enforcing the risk management related provisions and additional requirements related to data, advertising and transparency. Since 2015 it has also convened the EU Internet Forum which brings together relevant EU agencies such as EUROPOL and the Fundamental Rights agency, Member State authorities, industry actors, and other institutions, including the UN Office on Counter-terrorism to collaborate on illegal content. GIFCT and TAT have both been involved since their founding. The Forum was founded to address terrorist content, but since has also begun to tackle CSAM and trafficking.

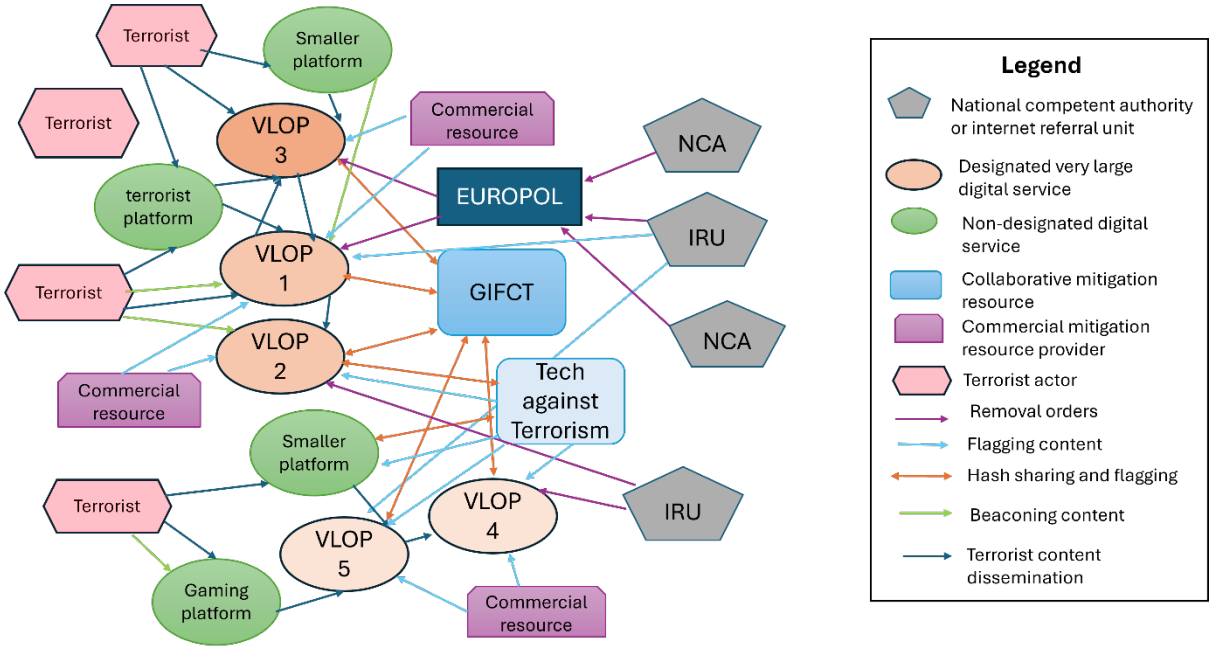
3.2 The ecosystem and its vulnerabilities

Key features of the ecosystem of interlinked actors and common resources are the significant role of national level authorities, a relatively limited set of sources for the identification of content, and an

[33] GIFCT (2024) Global Definitions of Terrorism https://def-frameworks.gifct.org/global-definitions-of-terrorism/
[34] Tech Against Terrorism, (10/2024) Inclusion Policy https://www.terrorismanalytics.org/policies/inclusion-policy

imbalance of capacity and access to resources between very large services and smaller ones. These can be seen as vulnerabilities in the ecosystem that introduce risks to the effective prevention of the dissemination of terrorist content and to the protection of freedom of expression. The assessment of how the content moderation systems of a VLOP that hosts content or the algorithmic systems generating a VLOSE's results, or either of their terms and conditions influence the risk of the dissemination of terrorist content[35] should consider how they are affected by these vulnerabilities. Figure 1 is a simplified illustration of these interlinkages, which include dissemination of content and various connections related to mitigation measures.

Figure 1 Simplified illustration of the actors and relationships involved in the dissemination of terrorist content and its mitigation



Source: the author

Interlinkages with other platforms and websites, especially smaller ones: There are two types of vulnerability here to consider. These stem from the evidence that a great deal of terrorist content is disseminated on smaller hosting services and websites, a phenomenon attributed to a lack of capacity or intention to prevent it. As Figure 1 illustrates, some smaller services are engaged in mitigation, while others are not and may even promote the dissemination of terrorist content. VLOPs, particularly social media and video-sharing platforms, and VLOSEs are often interlinked with these via the content disseminated and shared users.[36] The first consideration is how content coming from other platforms that are sources of terrorist content is handled by algorithmic systems and content moderation- and enabled by functionalities. One example is when a search result includes a video-sharing platform post that is terrorist content or a website that contains a video taken from a terrorist live-stream.

[35] These features are a reference to those noted in the DSA Article 34.2.

[36] See evidence and discussion in previous CERRE Report: Broughton Micova and Calef, 'Elements for Effective Systemic Risk Assessment under the DSA'.

The second consideration is how the service is being used to draw users to other services that are less moderated, unmoderated, or deliberately set up for terrorist purposes. This use of beacons, as mentioned above and illustrated with green arrows in Figure 1, relies on using borderline or coded content shared on VLOPs to attract users onto smaller online platforms or file-sharing services, or into groups on messaging services, and is particularly(though not exclusively) used by larger organised terrorist organisations.[37] There seems to be a lack of understanding of, and evidence on, the dynamics along these interlinkages. There are some efforts to make resources available to smaller services to help them remove content, but this appears to be unidirectional. The practice described by interviewees was that smaller services (those at all interested in combatting terrorist content) were very much on the receiving end of resources and tools, where offered, accepting identification of content unquestioningly and not feeding back into the resources or providing insight themselves. There was less clarity on the barriers to engagement.

Role of untransparent bodies in the identification of terrorist content and system testing: Because of the transparency and reporting obligations on service providers in the TERREG and the DSA, there is a certain amount of visibility into what VLOPs and VLOSEs are doing and how they are defining terrorist and borderline content. However, there are several actors involved in identifying terrorist content for which there is not potential for oversight, and this introduces vulnerability that stems from the definition problems discussed in Section 2. The NCAs, including IFUs, may be over or under inclusive in their removal orders and may be using flagging in parallel, as illustrated in Figure 1. There is potential for abuse of removal orders or their authoritative status as flaggers for political purposes. There is also a lack of transparency in the work of the commercial companies in terms of their inclusion criteria for the databases, insight, or red-team testing that they provide. GIFCT publishes annual transparency reports that includes information on the taxonomies and criteria used by its hash-database. TAT is transparent about its reliance on certain designation lists in its inclusion criteria. None of the commercial providers identified seemed to be transparent, nor do they seem to be included in the collaboration mechanisms where criteria, distinctions and principles are being discussed, such as GIFCT and the EU Internet Forum, each of which operates with some level of transparency. This introduces a challenge in terms of accountability for the effectiveness of mitigation measures and the extent to which they also avoid overly impinging on fundamental rights.

Insufficiently stable and accessible informational resources: Another vulnerability that can be seen in the system stems from the fact that the identification and removal of terrorist content requires a combination of automated and human intervention and relies on the ability to characterise pieces of content. GIFCT is a well-resourced collaboration to which major global tech companies contribute, and it operates a hash-database of terrorist content that is an important resource for stemming the dissemination of such content across multiple services. However, only the 32 providers who are members of GIFCT can access it. It has a mentorship programme to help services meet the criteria to join, but there is no evidence that it has instituted the tiered membership or removal process suggested in its last human rights audit.[38] TAT, which also maintains a database and provides notifications of terrorist content, focuses on serving smaller service providers, but it is a non-profit

[37] Elaborated in detail by CS Part 3, corroborated by CS Part 1 and CS Part 4.

[38] BSR, 'Human Rights Assessment: Global Internet Forum to Counter Terrorism.' (BSR, 2021), https://gifct.org/wp-content/uploads/2021/07/BSR_GIFCT_HRIA.pdf.

organisation that is funded by various pots of non-permanent public funding. Others in the system who provide resources are mostly commercial providers with commercial incentives. They may not be accessible to smaller providers and may leave the market at any time.

As one interviewee for this paper pointed out, for assessment and mitigation of the systemic risk of the dissemination of terrorist content "there needs to be some sense of being within a broader Tech sector or ecosystem and some sense of responsibility is also required, as there's no success really in isolation."[39] Another likened the effort of combatting terrorist content as a game of whack-a-mole because of the ways terrorists shift in the services and tactics they use.[40] Combatting this type of illegal content, and likely others as well, may require innovative pro-active approaches in addition to reactive measures on content. Pooling insight on patterns in user content-sharing behaviour, user migration, keywords and images among those engaged in prevention could enable this.[41] At the same time, failure to address terrorist content on smaller services increases the risk on VLOPS and VLOSEs, and failure to share information and pool knowledge among digital services and with other actors engaged in countering terrorism hampers mitigation. Systemic risk assessment should consider the vulnerabilities in the wider ecosystem and the ways that VLOP and VLOSE providers are contributing to exacerbating or reducing those vulnerabilities.

[39] CT Participant 1
[40] CT Participant 3
[41] This refers to insight on trends in anonymised or pseudonymised data in line with data protection rules and commitments to user privacy.

4. Benchmarks for risk management

From the understanding of the definitional issues and weaknesses in the wider ecosystem, we can derive two categories of benchmarks for the management of the systemic risk of dissemination of terrorist content. The first category covers what should be achieved in relation to the circulation of terrorist content on an individual VLOP or exposure to terrorist content via an individual VLOSE. These benchmarks should be used to measure the success of risk mitigation measures internal to each provider's service and user-base. The second category covers the contribution to prevention in the wider ecosystem that should be achieved – a kind of good citizenship marker – which is necessary because of the interlinkages and circulation patters discussed above. This section describes four broad benchmarks, two in each category, derived from international and EU law, declarations and commitments in the context of the UN, OSCE, the G20, and the Christchurch Call as well as the GIFCT criteria and input from interviewees.

4.1 Internal mitigation

The overarching aim of risk management, in terms of internal mitigation of the dissemination of terrorist content, should be to achieve the maximum possible elimination of terrorist content while also achieving maximum possible protection for freedom of expression and other fundamental rights. The need for efforts to combat terrorist content to incorporate safeguards for freedom of expression is mentioned in nearly every document examined for this paper. It was also a consistent message in the interview data. The DSA itself insists on "proportionate" as well as "effective" measures to mitigate risks implemented with "particular consideration" to their impact on fundamental rights.[42]

It is likely impossible to put a complete stop to attempts by terrorists and extremists encouraging terrorism to use digital services for their purposes in today's world. However, it is a reasonable benchmark to set, for the success of risk management, that users are not exposed to such content. This is therefore the first benchmark set out in Table 1 below. Such an end result is the aim of the TERREG, the Christchurch Call and other strategies and calls to combat terrorist content online. The ICCPR's article 20 legitimises efforts to limit such speech with its prohibition on incitement to hostility or violence.[43] There are two main aspects to limiting or eliminating exposure: the removal of content from platforms and ensuring that recommender systems and search results do not enable access to such content. The later includes the borderline content that can act as a gateway or facilitator of terrorist content, referred to in the DSA's Recital 84.

As discussed in Section 3, there is a whole ecosystem involved in efforts to stem the dissemination of terrorist content, so Table 1 below sets out some expectations for VLOPs and VLOSEs identified in the DSA, the TERREG, the interview data, and various international commitments. These can be seen as the elements of the contribution that these services should be making toward the mitigation of the systemic risk of the dissemination of terrorist content. One expectation raised by four of those interviewed is that service providers undertake thorough reflection and investigation into exactly how

[42] DSA Art 35.1

[43] The ICCPR Art 20 also prohibits propaganda for war and its exact wording on incitement refers to "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."

their services are being used for the dissemination of terrorist content. It was argued that understanding why particular functionalities, attributes, technical features and even business models might appeal to terrorists or be conducive to dissemination was crucial and that content moderation alone was not sufficient.[44]

The second benchmark shown in Table 1 sets the standard for what should be achieved in relation to fundamental rights, especially the rights to freedom of expression and assembly and the ability to engage in peaceful protest. Essential to protecting these rights is clarity and transparency about how terrorist content is defined and how actionable borderline content is determined. As was pointed out by one interviewee, definitions that refer to "dangerous individuals and organisations" or other vague terms disproportionately target historically oppressed groups.[45] At issue is whether the ways in which the designated services are determining purpose or intent and thus the illegality of the content are achieving the right balance with fundamental rights. Also at stake is how actioned borderline content is identified and treated, and whether adequate protections for fundamental rights are in place. Transparency about definitions and how lines are drawn can allow civil society organisations and those affected to challenge definitions.

The ability for users to appeal the actioning of content and accounts is also crucial to the protection of their fundamental rights. The expectations for VLOPs, primarily those for sharing user-generated content, and VLOSEs identified in Table 1 reflect the fact that users need to be able to know that their content has been removed or their website has been de-listed and why, and they need to have the ability to appeal against this action. This requirement for notification and explanation to the user is not absolute, however, in cases of terrorist content. According to TERREG's Article 11, competent authorities may determine that non-disclosure is required for up to 12 weeks for reasons of public security. According to its recital 34, the DSA's information requirements are without prejudice to this exception to disclosure obligations in TERREG and other EU laws covering specific types of illegal content and national criminal procedural law. Transparency and public reporting, on the scale and outcomes of appeals and on the types of complainants and their reasons for making complaints, is not just important for assessing the effectiveness of measures but can also reveal whether there are embedded biases in the definition or identification processes and whether there are any groups that are disproportionately affected by the removal of content.

[44] CS Part 1, CS Part 2, CS Part 3, DG Part 1
[45] DG Part 1

Table 1 Proposed benchmarks for the success of risk management related to internal mitigation of terrorist content with expectations for VLOPs and VLOSEs[46]

Benchmark	Source of benchmark	Expectations of VLOPs	Expectations of VLOSEs
1. Users are not exposed to terrorist content. - Terrorist content is identified swiftly and removed. - Search and recommender algorithms do not enable exposure to terrorist content.	Art 20 ICCPR; Christchurch Call Commitments; TERREG; G20 Osaka Leaders' statement on preventing exploitation of the internet for terrorism and violent extremism conducive to terrorism.	Providers fully understand how and why terrorists use their services' functionalities and technical features. Providers respond to removal notices from competent authorities in line with TERREG and to notices from others. Providers take measures both automated and with human intervention to identify and remove terrorist content and ensure the utmost accuracy of both. Borderline content is not amplified by recommender systems. Providers address out-linking identified to terrorist content and other beacons. Providers operate user-friendly notice & action mechanisms (DSA Art 16). Providers temporarily suspend users or repeatedly use the service to disseminate terrorist content (DSA Art 23)	Search results do not return the webpages or terrorist content hosted on other platforms. Sites are delisted promptly upon notification. Search results do not rank highly borderline sites and content.
2. Freedom of expression and rights to assembly and protest are protected.	Art 19, 21, 22 ICCPR; Art 9,10 & 11 ECHR	Providers maintain a clear definition of terrorist content that is easily accessible.	Providers maintain a clear definition of terrorist content easily accessible.

[46] A full list of sources with links is in Annex 1.

- Terrorist content is clearly defined. - Boundaries around borderline content are transparent and clear. - Appeal and redress mechanisms exist at all levels.	Christchurch call Commitments; GIFCT criteria; UN Global Counter Terrorism Strategy 8th Review; OSCE Charter on Preventing and Combating Terrorism and Ministerial Council Decision 07/06	Providers transparently establish and regularly review boundaries of borderline content. Services provide a statement of reasons (DSA Art 17) and notify the content creator when removing content (unless there is a public security exemption). Automated identification and removal is accompanied by human moderation, and regularly reviewed for accuracy. Providers operate an internal complaints-handling system (DSA 20).	Services provide a statement of reasons (DSA Art 17) and notify the content creator when de-listing content (unless there is a public security exemption). There is a clear appeals process for owners of sites that have been delisted.

4.2 Good citizenship and collaboration

In this second category are two benchmarks that reflect the interlinkages within the ecosystem and the need for cooperation and collaboration among services and other actors to manage the systemic risk of the dissemination of terrorist content. Cooperation among the providers of VLOPs and VLOSEs and other major providers, such as takes places among the members of GIFCT, is important to reaching these benchmarks. However, evidence from the interviews and from multiple reports from relevant actors examined for this paper indicates that cooperation among the largest services is not sufficient. Smaller platforms, law enforcement and other authorities, as well as civil society actors, need to be more involved.

The third benchmark, noted in Table 2 below, requires cooperation among these actors for the purpose of combatting the dissemination of terrorist content. Here the largest digital services, especially VLOPs, have the most information on content and behaviour, resources for identifying content, and expertise on mitigation mechanisms. However, as discussed above, many of the most intensively used services are not VLOPs and content on smaller services may make its way back onto VLOPs or lawful content on VLOPs may be used to direct users to terrorist content elsewhere. In order to mitigate the systemic risk of the dissemination of terrorist content, the providers of designated services will need to contribute to the elimination of terrorist content on other services as well.

On the one hand, as noted in the expectations listed in Table 2, this entails contributing as much as possible to increasing the capacity of smaller services to combat terrorist content on their services and engaging them in the creation of shared resources. The hash-database maintained by GIFCT is a good example of this kind of a resource. Such resources should be able to be used by smaller services to identify and remove terrorist content. Cooperation with smaller services providers can also help improve resources such as the hash-database if they contribute as well as receive. This can ensure that they are adequate to the systemic threat of terrorist content and not only to what is disseminated on the larger, usually well-moderated platforms, and can help with the identification of mistakes in hashing or classification that can impact freedom of expression if perpetuated.

On the other hand, the two benchmarks proposed in Table 2 cover contributions to efforts by law enforcement and other public authorities to investigate and take action against the dissemination of terrorist content and terrorism. This includes both efforts against terrorist groups and services that have no interest in combatting or may be designed specifically for the dissemination of terrorist content. Information held by VLOPs and VLOSEs can make significant contributions to such efforts, which are vital to the mitigation of the systemic risk. It was noted by interviewees and evident in the reporting of the services and other sources that there have been significant improvements by many VLOPs in removing known content, both benchmarks three and four reflect the need for proactive, insight driven cooperation to go after terrorist networks and other relevant malicious actors as well. As one interviewee commented, given the data possessed by the very large services, “with not much effort you can track and see how they [terrorist networks] evolve and are resilient to the more reactive takedown.”[47]

[47] DG Part 2

As listed in the expectation columns in Table 2, providers should have channels and systems for communicating with law enforcement or other competent authorities over specific instances of terrorist content and investigators will often need access to data from the platform about the content as evidence. For the wider prevention of the dissemination of terrorist content, and potentially even acts of terrorism, information from across multiple services and insight from content moderation efforts are essential. According to interviewees and other data sources examined for this paper, there is still a lack of standardised comparable reporting on incidence and actioning of terrorist content.[48] This inhibits the potential for learning from this data, discussed further in Section 5. Success in mitigating the systemic risk of the dissemination of terrorist content, is understanding how terrorists use different features, functionalities and types of services, which as one interviewee explained requires "information at scale."[49] VLOPs and VLOSEs, especially the former, hold a lot of this information and can make significant contributions to the wider efforts to combat the dissemination of terrorist content.

The live incident broadcast is a particular feature of terrorist content. While live-streaming is also implicated in the perpetration of child sexual exploitation and other illegal content, these are usually not intended to reach the most possible viewers. As mentioned before, while recruitment and radicalisation efforts might be conducted subtly, terrorists aim for their terrorist acts to receive as much attention as possible. Therefore, benchmark 3 for risk management in relation to terrorist content incorporates the requirement that systems are in place to collaboratively respond to such incidents. The DSA sets out a crisis protocol to be drawn up by the Commission, and a Content Incident Protocol is operated by GIFCT. It is important that VLOPs and VLOSEs participate in such mechanisms and that these crisis or incident responses take into consideration the interlinkages with smaller services.

[48] This included testing of the Transparency database established according to the DSA.
[49] CS Part 1

Table 2 Proposed benchmarks for contribution to the management of the systemic risk of dissemination of terrorist content with expectations for VLOPs and VLOSEs[50]

Benchmark	Source of benchmark	Expectations of VLOPs	Expectations of VLOSEs
3. There is cooperation among public authorities, digital service providers, and civil society to prevent the dissemination of terrorist content. - Knowledge, expertise and resources are shared, especially with smaller platforms. - There is regular, comparable and transparent reporting on measures. - Systems are in place to collaborate in relation to terrorist incidents that include all necessary actors.	Christchurch Call Commitments; Osaka Leaders' statement; UN Global Counter Terrorism Strategy 8th Review; OSCE Ministerial Council Decision 4/15	Providers systematically share knowledge, expertise, and resources with each other, smaller services and law enforcement, and engage with them in the development of resources. Providers cooperate with trusted flaggers. Providers cooperate with each other and other stakeholders to set accuracy metrics and error thresholds for content moderation. Providers report regularly on mitigation measures and cooperation systems in a manner that allows cross-platform comparison and learning. Providers engage with crisis protocol.	Providers systematically share knowledge, expertise, and resources with each other, smaller services and law enforcement. Providers report regularly on mitigation measures and cooperation systems in a manner that allows cross-platform comparison and learning. Providers engage with crisis protocol.
4. Terrorism, including acts and groups involved, can be prevented and investigated by relevant authorities.	TERREG, DSA	Providers establish channels and tools for communication with law enforcement and security services or engage with existing ones.	Providers establish channels and tools for communication with law enforcement and security

[50] A full list of sources with links is in Annex 1.

- Channels exist to provide information to authorities quickly and consistently. - Insight on the behaviour and tactics of those disseminating terrorist content is generated and made available to relevant authorities. - Evidence is protected and accessible to authorities.		Where there is suspicion of crime or threat life or safety, providers must notify law enforcement. Providers must maintain data that could be evidence for 6 months to allow investigation (TERREG), and otherwise comply with EU rules on electronic evidence.	services or engage in existing ones. Providers must maintain data that could be evidence for 6 months to allow investigation (TERREG), and otherwise comply with EU rules on electronic evidence

The benchmarks proposed in Tables 1 and 2 are not markers of compliance with the DSA for individual firms or services. They are benchmarks of success for in the mitigation of systemic risk of the dissemination of terrorist content, which all the evidence reviewed for this paper indicates is necessarily a collective endeavour. The eradication of terrorist content is not an achievable aim for the DSA, even combined with the TERREG and existing coordination efforts. However, the assessment and mitigation of the systemic risk of this type of illegal content as foreseen by the DSA can make a significant contribution. Systemic risk management goes well beyond the notice and take-down processes set out in the TERREG and other parts of the DSA.

As the benchmarks and expectations set out here highlight the importance of shared knowledge, collective insight, and cooperation on this type of illegal content. This requires risk assessments to be focused as much on each service's role in those as on its own ability to prevent the dissemination of terrorist content on its services. As one interviewee stated, success would be "If the risk assessment that they were conducting wasn't entirely inward looking but if it were actually more outward looking too."[51] The benchmarks set out here are also not only markers of success for the designated services, but of the risk management approach of the DSA to illegal content. They should be seen as the yardsticks against which to examine the outcomes of the assessment, mitigation, and reporting cycles that are instituted by the Act.

[51] CS Part 2

5. Metrics and Measurement

If the benchmarks are to be used to assess the effectiveness of risk management, there must be metrics and measures to indicate progress towards those benchmarks. Thanks largely to the reporting requirements of TERREG and the transparency requirements of the DSA, there is already some useful data available for evaluating attainment of the benchmarks, especially those related to internal mitigations.[52] The following are already required to be reported by digital service providers under each regulation.

TERREG (edited slightly for length)	DSA (edited slightly for length)
Article 7 (only for hosting services) • Information about measures in relation to the identification and removal of or disabling of access to terrorist content; • information about measures to address the reappearance online of material which has previously been removed or to which access has been disabled because it was considered to be terrorist content, in particular where automated tools have been used; • In response to removal orders, the number of items of terrorist content removed or to which access has been disabled and the number where the content was removed or access to which has not been disabled with the grounds for that decision; • the number and the outcome of complaints handled; • the number and the outcome of administrative or judicial review proceedings brought by the provider	**Article 15** • The number of orders received categorised by the type of illegal content, the Member State issuing the order, and the median time needed to inform the authority of its receipt, and to give effect to the order; • (only hosting services) the number of notices submitted categorised by the type of alleged illegal content concerned, the number of notices submitted by trusted flaggers, any action and whether on the basis of law or terms and conditions, the number of notices processed by using automated means and the median time needed for taking the action; • meaningful and comprehensible information about the content moderation engaged in at the providers' own initiative, including the use of automated tools, the measures taken to provide training and assistance to persons in charge of content moderation, the number and type of measures taken that affect the availability, visibility and accessibility of information, categorised by the type of illegal content or violation of the terms and conditions of the service provider, by the detection method and by the type of restriction applied; • the number of complaints received through the internal complaint-handling systems and, for providers of online platforms, the basis for those complaints, decisions taken and the median time needed for taking them and the number of reversals; • any use made of automated means for the purpose of content moderation, including a

[52] Some VLOP and VLOSE providers published some data voluntarily on government notices and removal actions before required by EU law and those in scope of Germany's NetzDG law also had reporting requirements under that law.

	qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error of the automated means used in fulfilling those purposes, and any safeguards applied. **Article 42 (Only VLOPs)** • the human resources that the provider of very large online platforms dedicates to content moderation in respect of the service offered in the Union, broken down by each applicable official language of the Member States, • the qualifications and linguistic expertise of the persons carrying out the activities referred to in point (a), as well as the training and support given to such staff; • the indicators of accuracy and related information broken down by each official language of the Member States.

The provisions of TERREG and Article 15 of the DSA apply to many more services than those designated as VLOPs and VLOSEs. This could offer potential for analysis of mitigation across the ecosystem, of the movement of terrorist activity, and of potential vulnerabilities. At the time of writing there was little to no detail in terms of the types and sources of actioned terrorist content. However, as of 4 November 2024, the Commission adopted a regulation that standardises the content and reporting periods for non-designated services and designated services under the DSA, so reporting in the coming years should be more conducive to such analysis.[53] The template attached to the new regulation includes a category code for reporting quantitatively on actioned terrorist content. The vast category of non-designated services includes online gaming services, where there are likely to be service specific types of data and particular concerns due to the live interactive formats. There may be a need for more nuance in the categories or understandings of the category of terrorist content in reporting.

Reports for the purposes of TERREG already provide metrics that can be used to assess the speed at which terrorist content is actioned, the volume and sources of removal orders, the existence of appeal mechanisms. Some of those examined for this study describe involvement with collaborative initiatives such as GIFCT and TAT and/or communication with law enforcement, but this is not consistent. Once services begin to use the transparency reporting templates adopted on 4 November, there will be much more granular detail on the reasons for and responses to removal orders. The additional reports required of VLOP providers in Article 42 also give an indication of the role that human moderation plays in relation to automated decisions and the accuracy of the automated systems as measured by the VLOP providers, which is are important indicators for benchmark 2.

[53] The regulation, templates and instructions can be found at https://digital-strategy.ec.europa.eu/en/library/implementing-regulation-laying-down-templates-concerning-transparency-reporting-obligations#:~:text=The%20Implementing%20Regulation%20standardises%20the,uniform%20reporting%20templates%20and%20periods.

In addition to the regular reports being generated in response to these provisions by the services in scope of the TERREG and the DSA, the Commission hosts the DSA Transparency Database to which services submit information on content moderation decisions on a daily basis. This data is exportable, API accessible and standardised. However, as of the time of writing there was no distinct statement of reasons category for terrorist content, only "illegal or harmful speech" and "violence". Some individual VLOP providers also have other API-based tools available, and all are mandated to have ad libraries.

Service providers are not the only ones producing relevant information. The DSA requires Digital Services Coordinators to generate reports on the removal orders that were issued by their Member State Authorities including the responses to those orders. The OECD's benchmarking project tracks policies and actions on terrorist content and violent extremism, providing important data on the evolution of definitions, terms, and transparency reporting practices. It, however, is limited to the 50 most popular global services and the 50 identified as most intensively used by terrorist groups (only some of which overlap in any given cycle).[54] GIFCT publishes annual reports with data on the accuracy of its hash database, the activation of its Incident Response Framework, and information requests from governments. These provide insight into cooperation among the services that are members of GIFCT in relation to benchmark 3 and information on the way terrorist content is being defined and identified for the hashing of individual pieces. The EU Internet Forum annual reports provide information on collaborative efforts within that forum to establish common understandings on emerging types of content and borderline content, as well as to improve communication systems with law enforcement and crisis responses. These reports from the Forum do not provide details on the contributions to this made by VLOP or VLOSE providers.

There is more data than ever before on the decisions made about content, actions taken and outcomes of appeals that speak to the expectations of the first two benchmarks. There is also evidence of cooperation and knowledge sharing as expected for benchmark 3 and communication with law enforcement as expected for benchmark 4. However, there are some specific gaps that will hinder a thorough evaluation of the contribution of VLOPs and VLOSEs to mitigating the systemic risk of terrorist content dissemination and the risk management approach to this type of illegal content overall.

The first and arguably most problematic gap, already indicated above, is the lack of standardisation and comparability of the data on actions taken on terrorist content and borderline content. The implementation on the recent regulation on transparency reporting by designated and non-designated services will go some way to rectify this. However, it seems there will still be gaps in both nuance and accuracy metrics. There is a further need for establishing specialised categories or nuanced elaborations of categories that are specific to the different types of services, as well as type or functionality specific metrics on accuracy. While accuracy targets for live or high-speed content might need to be different than for more static or slower disseminated content and privacy protection needs will vary across services, some standardisation and transparency on the setting of those targets would contribute to benchmarks 1,2 and 3. Cross-service analysis and collaboration among designated

[54] For the full list see the OECD's 2024 report https://www.oecd.org/en/publications/transparency-reporting-on-terrorist-and-violent-extremist-content-online_901cb8cf-en.html

services and non-designated services could be used to arrive at nuanced expectations for targets for actioning content and ways of reporting and tracking effectiveness.

The second is in metrics that would indicate success in mitigating exposure to terrorist content and borderline content. For example, we may have data on average speeds of removal, but it is also important to know how many people were exposed during the time that the content was available and how they came to be exposed; how quickly users can find or be presented with terrorist content or borderline content. An approach for assessing the effects of algorithmic measures – from the point of view of user experience – was piloted by researchers on behalf of the EU Internet Forum in 2023.[55] This kind of testing, done across services and in repetition systematically would be needed to see the extent to which each benchmark is being achieved, and indicate how much content is being missed. Also relevant to benchmarks 1 and 3 is data on the paths that content takes between services, for example, data on where/when it moves from other services to a VLOP or how/from where it moves from a VLOP that has removed it to other services. More data is needed on any measures VLOPs might have to deal with out-linking or beaconing to other services, as is data from the implementation of those measures that would contribute to insight on the nature and intensity of the use of such tactics.

Another area where there is a gap is information that would help to evaluate the ways that services' relationships with law enforcement and other competent authorities contribute to the mitigation of the systemic risk. There are some qualitative accounts in reports of those who participate in the EU Internet Forum and transparency from services and DSCs on the volume of removal orders and service responses to them. However there does not seem to be data available on law enforcement use of content data, such as on how long data is usually kept for law enforcement to access and how (and how often) they are using it. Such data could provide metrics for both benchmark 4 on enabling investigations and would be important for assessing benchmark 2 on achieving maximum protection for fundamental rights. This is clearly an area that merits more attention from researchers, regulators and civil society watchdogs, but at the moment there does not seem to be sufficient evidence upon which to base discussions or effective monitoring.

[55] European Union, *EU Internet Forum: Study on the Role and Effects of the Use of Algorithmic Amplification to Spread Terrorist, Violent Extremist and Borderline Content* – final report. October 2023.

6. Recommendations

This issue paper had two purposes: to arrive at benchmarks for evaluating the management of the systemic risk of the dissemination of terrorist content and to gather lessons that may be relevant for illegal content more generally.

It argued for benchmarks on internal mitigation that address both the exposure of users to content and the protection of fundamental rights. These go beyond basic targets for reacting to removal orders and actioning identified content. There is a need for each VLOP or VLOSE provider to thoroughly interrogate how their services are being used and what functionalities make them appealing to terrorism-related bad actors, including linkages with other services. More nuanced targets or expectations specific to certain functionalities or interlinkages (such as with private messaging or gaming platforms) are needed. At the same time inclusive discussions that bring together service providers, various prevention-focused actors, fundamental rights experts and groups affected by mitigation measures are needed on definitions and boundaries of borderline content and on the relationships with law enforcement and state security services in order to ensure protection of fundamental rights.

It was also argued that the management of systemic risk in this area should be evaluated against benchmarks for collaboration and contribution to mitigating the dissemination of terrorist content in the wider ecosystem. Combatting the dissemination of terrorist content, and likely most types of illegal content, requires cooperation among services and with other actors, such as law enforcement and various types of third-party flaggers. Fundamental to these benchmarks are knowledge exchange, sharing of resources and insight development, all within the limits allowed by data protection rules. In relation to terrorist content, there were some good examples of cooperation happening. However, the investigation identified instability in the resources for smaller services and intransparency about the roles of various actors, especially law enforcement and commercial suppliers of content moderation resources.

Several lessons from this examination of the issue of terrorist content are applicable for wider efforts on illegal content, including the need to balance removal or exposure prevention targets with fundamental rights targets and the importance of collaboration and knowledge sharing. There are often overlaps between terrorist content and other categories of illegal content- and borderline content is a grey area across multiple categories. Law enforcement or other authorities are also involved in efforts to combat various types of illegal content.

Most centrally, the case of terrorist content indicates that an effective approach to managing the systemic risk of illegal content dissemination requires the structured involvement of several overlapping sets of actors and the development or reinforcement of mechanisms for coordination and communication among these actors. This paper makes three generalisable recommendations to avoid duplication across the various kinds of illegal content and ensure adequate protection of fundamental rights.

1. There is a need for inclusive mechanisms that involve VLOP and VLOSE providers, law enforcement, regulatory authorities, experts, and civil society groups – especially those representing groups affected by mitigation measures – to facilitate dynamic, transparent

understandings of borderline content and how it is treated. These should join up with discussions on hate speech and disinformation for which there are already relatively inclusive collaborative efforts at the EU level.

2. Evaluation of the assessment and management of systemic risk of illegal content dissemination should consider how larger services are engaging with smaller services, especially in boosting their capacity to contribute to mitigation, and it should make use of nuanced and transparent standards for metrics on actioning content including accuracy targets or error thresholds.
3. Designated service providers and Member State authorities need to facilitate accessible and detailed data about their channels of communication and the use of data held by services and/or law enforcement for the purpose of investigation to enable better understanding and oversight of the role of law enforcement and other authorities.

Terrorist content may be less abundant than other types of illegal content, but it can have significant and widespread impact. It does merit special attention for that reason, and because it is where the interaction between harm-specific legislation (in addition to criminal law) and the DSA can be seen. This issue paper therefore makes two specific recommendations related to the achievement of the benchmarks related to terrorist content.

4. When the first round of risk assessments and their audits are publicly available to all the DSCs, the designated services, academics and the wider stakeholder community, analysis should be conducted to look across assessments to understand and evaluate the roles that different actors play in the mitigation of risk. In relation to terrorist content five areas of analysis are important starting points.

    - Analysis of the role played by institutionalised cooperation, the EU Internet Forum, and GIFCT, in mitigation
    - Analysis of the extent to which mitigation depends on commercial third parties and the level and type of oversight or transparency that is in place, if any, in relation to their contribution
    - Analysis of how borderline terrorist content is being determined and handled across platforms, particularly in relation to the tactic of beaconing and the processes of radicalisation and incitement
    - Analysis of the extent to which there is evidence of VLOP and VLOSE providers engaging with smaller services and the form that such engagement takes
    - Analysis of the extent to which each provider's assessment of risk considers their service's contribution to a cycle of sharing insight, expanding collective understanding of the problem, and investment in improving mitigations

    If there is not sufficient evidence in the risk assessments to conduct analysis in these areas, then they should serve as guides for improvements in future rounds of risk assessment.

5. Aside from the insight that may be gained from the risk assessments by designated services, there is a need to build on the standardisation of reporting by smaller services now underway. More nuanced, yet standardised data would enable better comparative analysis to improve

mitigation and collaboration, so the reporting by non-designated services should achieve more detail on measures taken on terrorist content and responses to those measures. Those evaluating the management of the systemic risk of dissemination of terrorist content, such as the Digital Services Board, the EU Internet Forum, and civil society groups, will need to look at what is happening across the wider ecosystem and be able to take this information into account.

cerre Centre on Regulation in Europe

Avenue Louise 475 (box 10)
1050 Brussels, Belgium
+32 2 230 83 60
info@cerre.eu
www.cerre.eu

Centre on Regulation in Europe (CERRE)
CERRE Think Tank