EJLT European Journal of Law and Technology

Transparent, Rapid and Contextualised? Comparing Content Moderation Requirements in Emerging EU Regulation

Therese Enarsson*

Abstract

The European Union is making unprecedented efforts to address illegal and harmful online content through regulation, particularly targeting very large online platforms (VLOPs). This article examines how recent EU instruments – starting with the Digital Services Act (DSA) – assign responsibility for efficient and rapid content moderation while allowing for contextualised decision-making. This has implications for both automated and human moderation. The analysis shows that advanced moderation is effectively mandated, requiring a mix of automation and human oversight, though the explicit need for human contextual input is rarely stated. Instead, VLOPs are primarily tasked with ensuring transparency in moderation processes and decisions. As a result, responsibility is partly shifted to users, who must take certain actions, such as requesting human review of automated decisions. The article also highlights the complex interaction between the DSA and other specialised regulations and self-regulatory measures on hate speech, terrorism-related content and disinformation – creating a dense regulatory landscape that warrants further study.

Keywords: Digital Services Act, content moderation, very large online platforms, hate speech, disinformation, terrorist content

* Department of Law, Umeå University. This work was supported by the Swedish Research Council under Grant number 2020-02278.

1. Introduction

1.1 Regulating Online Content Moderation

Online content moderation has been described as 'one of the tech-world's best kept secrets'[1] as it has largely flown under the radar and has only in the past decade truly drawn public attention. This often-invisible practice is intended to keep platform content safe and relevant for both users and the platform itself. It involves identifying and reviewing harmful content and, if necessary, deleting or restricting access to such content or banning users. While the concept of content moderation is not new – long predating major online platforms, news media providers and even the internet – the vast volume of content that can now be shared in an instant has fundamentally altered the field. This development has prompted legal questions regarding how to address illegal or abusive content online.

In response, we have seen unprecedented efforts through European legislation targeting the online sphere. The clearest example is the landmark Digital Services Act (DSA), which builds upon and reforms the Electronic Commerce Directive (2000/31/EC). The DSA came into force in 2023 for very large online platforms (VLOPs), and in February 2024 for all online platforms operating in the European market.[2] With this, the European Union (EU) became the first jurisdiction to set standards for the responsibilities of VLOPs online. By increasing the accountability of those providing arenas for information sharing – particularly VLOPs – the EU aims to create safer, more transparent online spaces for users.[3] A key part of this is requiring VLOPs to protect users' fundamental rights during and throughout the content moderation process.[4]

By increasing the responsibility of VLOPs, their role in countering illegal content online is being clearly defined. This shift is a consequence of the European Commission's decade-long strategy to address not only illegal but also harmful (though not necessarily illegal) content online. In addition to the DSA, one such initiative is the

---

[1] Ysabel Gerrard, 'Social Media Moderation', in Devan Rosen (ed), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural effects of Social Media* (Routledge 2022) 77–95.

[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) 2022 (OJ L). VLOPs are here defined in accordance with Art 33(1) of the DSA, as online platforms with more than 45 million users. (The same amount of average monthly users defines a very large online search engine.)

[3] VLOPs will be the focus of this article, and the term will be used throughout, however 'platforms', 'service providers' and 'intermediates' are also used in this article's terminology, as well as in the regulatory framework studied. When used it should be understood as something that *is* applicable for VLOPs, even if it could sometimes be applied toward smaller platforms as well.

[4] Arts 34 and 35 DSA. See also, *inter alia*, Recitals 3 and 84.

Strengthened Code of Practice on Disinformation (the Code of Practice),[5] a self- and co-regulatory instrument and agreement between VLOPs, that clarifies their responsibility to protect users from disinformation.[6] The Code of Practice also became a code of conduct under the DSA in July 2025.[7] Another EU instrument, specifically targeting terrorist content – the Regulation on addressing the dissemination of terrorist content online (TERREG) – came into force in 2022.[8] Taken together, these actions reflect the EU's growing intention to regulate the spread of certain types of content on social media, placing the most extensive obligations on VLOPs to assist in these ambitions.[9]

These instruments were not the EU's first steps toward increasing VLOP responsibilities. A key precursor was the EU Code of Conduct on Countering Illegal Hate Speech Online ('the Code of Conduct'). This agreement was unique as it was the first instance where the Commission reached a voluntary agreement with VLOPs to moderate platform content.[10] Beyond its core objective of countering hate speech and terrorist propaganda, a key element of the Code of Conduct is that signatory VLOPs committed to reviewing, removing and reducing access to reported hate speech as quickly as possible – typically within 24 hours – for most such reports.[11]

Through the DSA, the Code of Practice, TERREG and the Code of Conduct on Hate Speech, a complex web of legally binding and voluntary requirements has emerged. Depending on the type of content, one or more of these instruments may apply, each imposing different timelines for moderation. These timelines implicitly affect the feasibility of using automated moderation as well as human moderation or oversight. Instruments like the Code of Conduct, with its 24-hour removal expectation, implicitly

---

[5] European Commission, *The Strengthened Code of Practice on Disinformation* (2022) <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation>. This is an updated version from 2022 of the EU Code of Practice on Disinformation (Code) from 2018.

[6] European Commission, *Tackling Online Disinformation: A European Approach*, Communication COM (2018) 236 final.

[7] European Commission, *The Code of Conduct on Disinformation* <https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation> accessed 7 March 2025.

[8] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online [2021] OJ L172/79.

[9] The responsibilities under the DSA can be described as an asymmetrical pyramid, where the obligations on different types of online intermediaries vary depending on their size and impact, but also the type of services provided. The smallest platforms are therefore exempted from most of the obligations targeting VLOPs, for example. See for example Miriam C Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation' (2021) 12 *JIPITEC* 361 p. 363 and Martin Eifert and others, 'Taming the Giants: The DMA/DSA Package' (2021) 58 *Common Market Law Review*.

[10] European Commission, 'The EU Code of Conduct on Countering Illegal Hate Speech Online' <https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 7 March 2025.

[11] Ibid.

require automated moderation technologies to process vast amounts of content in short timeframes. Others, like TERREG, explicitly emphasise the importance of human oversight to contextualise content and guard against over-moderation.[12]

To understand the emerging responsibilities for VLOPs, there is a need to map the relevant content moderation requirements across these instruments before comparing and analysing their implications in relation to speed and contextualisation.

This article aims to contribute to understanding how recent EU instruments regulate VLOPs' responsibilities for effective content moderation, specifically, demands that VLOPs act rapidly (indicating automation) and make contextualised decisions (indicating human involvement). It does so by mapping and comparing moderation requirements and relevant interpretive standards in the evolving EU regulatory landscape targeting VLOPs. The focus is on the DSA,[13] with analysis and comparison of its moderation requirements against those in the Code of Conduct, TERREG and the Code of Practice on Disinformation. This analysis will address the extent to which, and in which contexts, the growing regulatory framework directly or indirectly mandates automated decisions or human moderation to ensure contextual sensitivity.

The complexity of these overlapping instruments should not be underestimated. While the DSA attempts to comprehensively regulate platforms, its legal obligations are often goal-oriented and lack specificity. Frequent cross-references to content-specific instruments imply that their requirements may influence DSA interpretations. At the same time, all instruments emphasise interpretation in line with fundamental rights, while leaving discretion for platforms to self-regulate. Case law remains limited, and while the DSA imposes strict obligations with potential penalties, clear guidance from the Court of Justice of the European Union (CJEU) is lacking.

Accordingly, this study employs a qualitative legal analysis and comparison of these regulatory instruments and, where relevant, their *travaux préparatoires*. The textual analysis will focus on how VLOP obligations are constructed – particularly regarding automation, human moderation or oversight, and what moderation systems are explicitly required. Policy documents and, where available, legal proceedings will be used to provide further context for these instruments and the regulatory ambitions underpinning them. Though direct case law remains scarce, legal scholarship will be used to support interpretations and add necessary context.

The article is structured in three parts. The remainder of section 1 provides a background on online content moderation and the need to balance speed with

---

[12] Art 5(3) Regulation (EU) 2021/784 (n 8). This is further developed under section 2.2.3 below.
[13] Of course, it should be pointed out that intermediary liability for VLOPs and other large platforms to moderate certain content can also be affected by other regulation, like the European Convention on Human Rights. As stated, the main focal point of this article is however not the matter of liability, but how the responsibilities appointed to VLOPs are framed in new regulatory instruments.

contextualisation. Section 2 introduces freedom of expression as a conceptual backdrop to the evolving regulatory landscape. This is followed by a mapping and analysis of the relevant instruments, beginning with the DSA, then the Code of Conduct and TERREG, and finally the Code of Practice on Disinformation. Section 3 presents a final analysis in light of the research questions and draws overall conclusions.

1.2 Online Content Moderation – Keep, Remove, or Hide Content?

It is important to note that not all content moderation involves a binary choice of whether to delete content or ban users. Other forms of moderation, such as reduction, are also common on platforms. 'Reduction' refers to the restriction of the visibility or reach of content deemed problematic by the platform, though not sufficiently problematic to warrant removal under its guidelines.[14] One way of describing this is 'shadow banning,' since its effects – reduced visibility and reach – are similar to those of an actual ban. However, the process is often invisible and unclear (or even unknown) to the user. Shadow banning can be understood in different ways. One definition is the complete restriction of a user's visibility and reach without notification, so that the user continues to perceive their communication with the platform and others as functional, even though none of their content actually reaches others.[15] Another interpretation of shadow banning, however, is essentially synonymous with reduction as described above.

That said, content moderation also frequently includes banning users or deleting or limiting access to material, regardless of whether it contains content that others may find disturbing. Detecting potentially illegal or undesirable content is, of course, essential to enable platforms to make informed decisions about what to remove, reduce or retain. This process typically involves a combination of methods, such as automated systems that flag potentially illegal content based on certain phrases or known images (particularly in relation to child sexual exploitation), and human moderators who assess content based on both its relevance to the platform and user safety. Notice-and-action mechanisms are also widely used by VLOPs and are now mandatory under the DSA.[16] These systems must offer users electronic means of notifying platforms of potentially illegal content or content that violates platform guidelines. As a result, platforms – and, where relevant, their moderators – must not only follow company policies and internal guidelines, but also possess extensive knowledge of current national and European legislation and regulation.

---

[14] Tarleton Gillespie, 'Do Not Recommend? Reduction as a Form of Content Moderation' (2022) 8 *Social Media + Society* 1.

[15] Paddy Leerssen, 'An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation' (2023) 48 *Computer Law & Security Review* 105790. This form of blocking is also the definition used in the DSA (Recital 55).

[16] Art 16 DSA; Pieter Wolters and Raphaël Gellert, 'Towards a Better Notice and Action Mechanism in the DSA' (2023) 13 *JIPITEC* 403.

As previously noted, moderation processes involve a balance between automated systems and human moderators. Automated systems can analyse significantly larger volumes of content, and at much higher speed, than any human could. The sheer scale of content on VLOPs is often cited as the justification for using automated moderation.[17] In addition, automation can ease the emotional burden placed on human moderators, who are routinely exposed to disturbing material.[18]

However, what humans lack in speed or bandwidth, they compensate for in their ability to make complex, contextualised decisions. Human moderators can also help counter the inherent biases of automated systems.[19] Moreover, humans are better at interpreting aspects such as underlying intent, cultural nuance, idiomatic expressions and irony. A lack of such contextual understanding can result in infringements on freedom of expression through overly aggressive moderation. Excessive content moderation may produce a chilling effect, discouraging individuals from participating in political discourse due to a real or perceived sense of censorship. It may also dissuade individuals from expressing their religious beliefs, revealing their sexual orientation, or exploring their gender identity.[20]

There are reported instances of specific communities being silenced due to misunderstandings by automated systems. For example, content created by members of the drag queen community has been removed because of the misinterpretation of certain phrases.[21] Content posted by trans individuals is

---

[17] Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale' (2020) 7 *Big Data & Society* 2053951720943234.

[18] Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press 2019) 162–164; 'Ex-Facebook Moderator in Kenya Sues over Working Conditions' *The Guardian* (10 May 2022) <https://www.theguardian.com/technology/2022/may/10/ex-facebook-moderator-in-kenya-sues-over-working-conditions> accessed 7 March 2025; Adi Robertson, 'TikTok Moderators Say They Were Shown Child Sexual Abuse Videos during Training' The Verge (5 August 2022) <https://www.theverge.com/2022/8/5/23294017/tiktok-teleperformance-employees-shown-csam-moderation-report> accessed 7 March 2025; 'The Human Cost of Online Content Moderation' *Harvard Journal of Law & Technology* (2 March 2018) <https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation> accessed 7 March 2025.

[19] European Union Agency for Fundamental Rights (FRA), 'Online Content Moderation Current Challenges in Detecting Hate Speech' (2023). The human vs. automated detection is discussed in relation to detection of hateful speech, see 13–14.

[20] Jennifer Cobbe, 'Algorithmic Censorship by Social Platforms: Power and Resistance' (2021) 34 *Philosophy & Technology* 739; Greyson K Young, 'How Much Is Too Much: The Difficulties of Social Media Content Moderation' (2022) 31 *Information & Communications Technology Law* 1.

[21] Thiago Dias Oliva, Dennys Marcelo Antonialli and Alessandra Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 *Sexuality & Culture* 700; Maarten Sap and others, 'The Risk of Racial Bias in Hate Speech Detection', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2019) <https://www.aclweb.org/anthology/P19-1163> accessed 7 March 2025.

reportedly more likely to be classified as 'adult' under platform guidelines and thus automatically removed, absent any contextual understanding.[22] Additionally, people of colour and other minority groups have reported that their content is more heavily moderated than that of white or majority populations, suggesting that automated systems disproportionately protect dominant groups.[23] Conversely, the lack of contextual awareness in automated systems may also allow problematic content to remain online when it should be flagged or removed. For instance, automated tools may fail to keep pace with evolving slurs or symbols used in extremist communities. A well-documented example is the use of the number '88' (a coded reference to 'Heil Hitler') to avoid detection.[24]

Therefore, a combination of automated and human moderation may be necessary to meet legal requirements: on the one hand, to handle the vast volume of potentially abusive content in a clear, transparent, and efficient manner; and on the other, to do so without unduly infringing users' fundamental rights.[25]

The complexity of online content moderation should not be underestimated. As we will see, the challenges outlined in this section are reflected in the regulatory approaches to online platforms examined in the sections that follow.

2. The Emerging Legal Framework

2.1 Content Moderation against a Backdrop of Freedom of Expression

Providing more than a brief background to the right to freedom of expression online is beyond the scope and ambition of this article. However, it is important to note that the overall objectives and underlying principles of this right serve as a significant

---

[22] The content could, for example, include photos of surgical procedures. There have also been reports of accounts of trans people being removed for violating the 'real name' policy of Facebook. Oliver L Haimson, Kaniel Delmonaco, Peipei Nie and Andrea Wegner, 'Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas' (2021) 5 *PACM HCI, CSCW2,* Article 466, 1–35. <http://deepblue.lib.umich.edu/handle/2027.42/169587> accessed 7 March 2025.

[23] Haimson et al (n 22); see also Elizabeth Dwoskin, Nitasha Tiku and Heather Kelly, 'Facebook to Start Policing Anti-Black Hate Speech More Aggressively than Anti-White Comments, Documents Show' *Washington Post* (3 December 2020) <https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech> accessed 7 March 2025.

[24] See Andrew Kersley, 'The One Problem with AI Content Moderation? It Doesn't Work' Computer Weekly (7 February 2023) <https://www.computerweekly.com/feature/The-one-problem-with-AI-content-moderation-It-doesnt-work> accessed 7 March 2025.

[25] Therese Enarsson, Lena Enqvist and Markus Naarttijärvi, 'Approaching the Human in the Loop – Legal Perspectives on Hybrid Human/Algorithmic Decision-Making in Three Contexts' (2022) 31 *Information & Communications Technology Law* 123.

backdrop to the instruments examined. As such, a few key points summarising relevant principles will be outlined here.

In the Charter of Fundamental Rights of the European Union ('the Charter'), the right to freedom of expression is primarily articulated in Article 11, which states in Article 11(1):

> Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.

When regulating the responsibilities of online platforms, the CJEU has established that such regulations must strike a balance between competing interests – such as users' rights to freedom of expression and access to information under Article 11, and platforms' rights under Article 16 of the Charter to conduct a business.[26] The CJEU has also emphasised that any measures taken against specific content must be narrowly tailored so as not to unnecessarily infringe users' rights to information.[27]

In the landmark case *Eva Glawischnig-Piesczek v Facebook Ireland Limited*,[28] the CJEU further held that platforms such as Facebook can be required to remove illegal content uploaded to their services, even without prior notification, as well as content that is 'identical' or sufficiently similar in meaning to content already deemed illegal. This suggests that automated tools are to be used in identifying such content, and may even imply a degree of monitoring by platforms.[29] How this ruling is to be reconciled with the protection of fundamental rights remains unclear and potentially challenging.[30]

In the context of online content dissemination, the dual nature of the right to freedom of expression – to both receive and impart information – is particularly significant. Given their dominant market position, VLOPs can affect not only individual users

---

[26] Giancarlo Frosio and Christophe Geiger, 'Taking Fundamental Rights Seriously in the Digital Services Act's Platform Liability Regime' (2023) 29 *European Law Journal* 31, 35–36. As an example, this has been discussed in intellectual property cases from the CJEU concerning the balancing of interests when platforms are used to disseminate material infringing a third parties copyright, and the use of automatic filtering or review by platforms on users' uploaded content. Case C-314/12, *UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH and Wega Filmproduktionsgesellschaft mbH* [2014] EU:C:2014:192, para 47; Case C-401/19, *Republic of Poland v European Parliament and Council of the European Union* [2022] EU:C:2022:297, para 70.

[27] Case C-314/12 (n 26), para 56.

[28] Case C-18/18, *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] EU:C:2019:821, paras 39, 41, 45 and 46. See also Austrian supreme court in *Die Grünen v. Facebook Ireland Limited* [2020] Case 6 Ob 195/19y, and The Regional Court (Landgericht) of Frankfurt am Main, Germany in *Blume v Twitter* [2022] Case 2-03 O 325/22.

[29] Therese Enarsson, 'Navigating Hate Speech and Content Moderation under the DSA: Insights from ECtHR Case Law' (2024) 33 *Information & Communications Technology Law* 384, 392–393.

[30] Daphne Keller, 'Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling' [2020] 69 *GRUR International* 616, 623.

seeking to share specific content but also the broader public attempting to access information more generally. The ability to receive and communicate information is fundamental to participation in societal discourse, which in highly digitalised societies often takes place online.[31] At the same time, protecting users from illegal and potentially harmful content, such as hate speech and terrorist propaganda, is equally essential. Accordingly, the extent to which emerging regulations impact what content is permitted – or when and how it is subject to moderation – has implications for the protection of both users and society as a whole.

One overarching challenge is that the rights to freedom of expression and information are highly contextual. Expressions that may legitimately be restricted in one setting may be protected in another.[32] Therefore, constructing proportionate content moderation requirements must account for this contextual nature and the associated need for individualised assessments. This principle is also reflected in the European Convention on Human Rights (ECHR). The rights enshrined in the ECHR constitute general principles of EU law, as recognised in Article 53 of the Charter and Article 6(3) of the Treaty on European Union. The extensive case law of the European Court of Human Rights (ECtHR) interpreting Article 10 ECHR demonstrates that a nuanced, contextualised approach to freedom of expression is essential.[33] Such an approach allows for the restriction of harmful speech while avoiding a chilling effect on legitimate expressions. As noted, a detailed analysis of freedom of expression case law from the CJEU or ECtHR is beyond the scope of this article. Nonetheless, an awareness of this broader legal context is crucial, as it is likely to inform the interpretation of emerging EU legal requirements.[34] These requirements often demand context-sensitive assessments, rather than blanket restrictions on particular types of expression.

2.2 Digital Services Act

2.2.1 General Requirements Relating to Content Moderation

---

[31] Anni Carlsson, *Constitutional Protection of Freedom of Expression in the Age of Social Media : A Comparative Study* (Uppsala University 2024) 23–24.

[32] See Therese Enarsson and Simon Lindgren, 'Free Speech or Hate Speech? A Legal Analysis of the Discourse about Roma on Twitter' (2019) 28 *Information & Communications Technology Law* 1; Therese Enarsson and Markus Naarttijärvi, 'Is It All Part of the Game? Victim Differentiation and the Normative Protection of Victims of Online Antagonism under the European Convention on Human Rights' (2016) 22 *International Review of Victimology* 123.

[33] Again, to provide an in-depth analysis of the substantial case law from the ECtHR is beyond the ambition of this article, but see for instance *Handyside v the United Kingdom* (1976) App no 5493/72 (ECtHR); *Jersild v Denmark* (1994) App no 15890/89 (ECtHR); *Vejdeland and Others v Sweden* (2012) App no 1813/07 (ECtHR); *Lilliendahl v Iceland* (dec) (2020) App no 29297/18 (ECtHR); Hannes Cannie and Dirk Voorhoof, 'The Abuse Clause and Freedom of Expression in the European Human Rights Convention: An Added Value for Democracy and Human Rights Protection?' (2011) 29 *Netherlands Quarterly of Human Rights* 54.

[34] Giancarlo Frosio and Christophe Geiger, 'Taking Fundamental Rights Seriously in the Digital Services Act's Platform Liability Regime' (2023) 29 *European Law Journal* 31, 43.

Since the DSA came into force, it has introduced several changes to the regulation of online content moderation. The overarching goal of the Act is to increase transparency and user safety online by imposing obligations on all intermediary services, regardless of size, with additional duties placed on hosting services – particularly social media platforms and VLOPs.[35] The DSA addresses moderation broadly, including both removal or non-removal and the restriction of content visibility:

> '[C]ontent moderation' means the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account.[36]

In 2023, the EU further specified which platforms are considered large enough to be subject to the DSA's strictest regulatory requirements. Of the 19 designated services, 17 are VLOPs, and the remaining two are very large online search engines (VLOSEs).[37] The list of VLOPs didn't offer many surprises and included platforms like TikTok, Meta/Facebook, Google, Instagram and Twitter/X.[38] Each of these platforms reaches at least 45 million monthly active users and is therefore subject to a more stringent set of obligations under the DSA. For example, VLOPs must publish annual reports detailing their content moderation activities and how they enforce their terms and conditions.[39]

Other general obligations for VLOPs concerning content moderation are also worth noting. VLOPs are required to structure their infrastructure and mitigation systems to address the risks posed by the spread of illegal content online. This obligation is outlined in Articles 34 and 35 of the DSA, as well as in Recital 79, which emphasises

---

[35] Recital 41 DSA.
[36] Art 3(t) DSA.
[37] The definition of very large online platforms and very large online search engines ca be found in Art 33 DSA. See also Recital 76 ibid. where this is discussed in relation to the responsibility of these very large platforms. For a detailed list, see European Commission 'DSA: Very Large Online Platforms and Search Engines' <https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413> accessed 7 March 2025.
[38] 'DSA: Very Large Online Platforms and Search Engines' (n 37); this is specified in Recital 16 DSA. Also see Recitals 17, 18, 20 and 21 for what the platforms must live up to in order to benefit from this exemption.
[39] Art 15(2) and Recital 49DSA; European Commission 'Commission Recommendation of 6 May 2003 Concerning the Definition of Micro, Small and Medium-Sized Enterprises (Text with EEA Relevance) (Notified under Document Number C(2003) 1422)', (2003) 36 <http://data.europa.eu/eli/reco/2003/361/oj/eng> accessed 7 March 2025.

that limiting systemic risks and respecting fundamental rights is a societal responsibility of VLOPs. This requires VLOPs to balance various internal and external interests, employing all available means to counteract risks – particularly with regard to the impact on freedom of expression. This is to be weighed against the financial means and usability of the platform.[40] One factor under scrutiny is content moderation itself and how effectively it manages the dissemination of illegal content through the service.[41]

Understanding the general moderation responsibilities under the DSA begins with mapping out the relevant requirements, starting with Article 14, which concerns the use of terms and conditions. Article 14 makes it clear that VLOPs must explain how content moderation is conducted on their platforms, 'including algorithmic decision-making and human review', and must clarify for users how complaint mechanisms work.[42] The Article also states that VLOPs have a responsibility to act in a 'diligent, objective and proportionate manner' when restricting users on their platform and must take into account users' fundamental rights as expressed in the Charter. Article 14(4) specifically cites freedom of expression and media pluralism as such rights.

Article 14 can therefore be understood as requiring content moderation to be both efficient and balanced. It permits both automated and human moderation, but it also stresses the importance of transparency, respect for users' fundamental rights and the usability of complaint mechanisms. However, Article 14 focuses on communication with users and does not itself *require* human review to ensure the protection of fundamental rights or provide contextual understanding to moderation decisions. Moreover, it offers little guidance on how moderation should be conducted or on the use of automated systems.

Further insight into the DSA's expectations can be found in Article 15. This provision requires VLOPs to publish annual reports describing their use of automated content moderation, including assessments of accuracy and error rates.[43] This highlights the importance of understanding the capabilities and limitations of such systems. Like Article 14, however, Article 15 does not prescribe a specific moderation process. It does require platforms to report on the types of moderation used, whether automated tools were employed, and whether any technologies were used to restrict the visibility or accessibility of content – whether users were providing or receiving information. Platforms must also disclose the types of violations detected, the detection methods, and the forms of restriction applied. Notably, Article 15 also refers to human moderators, requiring platforms to provide information about the

[40] Recital 86 DSA.
[41] Art 34(2)(b) DSA.
[42] DSA. When an intermediary service targets children primarily, they must also take that into consideration in the presentation of the information so that it is easily understand, see Art 14(3) DSA.
[43] Art 15(1)(e) DSA.

training and support given to those responsible for moderation.[44] However, the Article does not explicitly mandate that such training or support be provided.

Similarly, under Article 17, VLOPs must inform users – upon request – about the type of moderation applied to their content.[45] For example, platforms must disclose whether content was removed, demoted or restricted, and on what basis. If relevant, they must also state whether automated tools were involved in the decision. Again, the DSA underscores the importance of transparency, particularly in enabling users and external actors to protect their interests. Still, it does not require the use of either automated or human moderation by default.

However, some provisions do relate to user access to human review. Recital 58 states that services should provide internal complaint systems to ensure that content is not wrongfully removed. Where automated systems are used, such complaint mechanisms should include human review. While this may offer an important safeguard for fundamental rights, it does not guarantee human review for each user. The DSA only requires human review when users are notified of a moderation decision and choose to contest it.[46]

Thus, the general requirements for content moderation under the DSA allow for both automated and human moderation. However, the DSA does not mandate the use of human moderators to provide contextual insight unless a user challenges a decision. The most consistent requirement is transparency regarding the methods and outcomes of moderation, which plays a key role in protecting user rights.

So far, we have addressed the overarching requirements relating to general moderation responsibilities. However, when it comes to illegal activities, more specific rules may apply.

2.2.2 Content Moderation and Illegal Content – the DSA Meets Content-specific Instruments

The provisions of the DSA that are more specific regarding content moderation often refer to other guidelines or instruments. A clear example is the previously mentioned

[44] Art 15(1)(c) DSA.
[45] Art 17(2)(a), 17(3)(a)–(c) DSA.
[46] This corresponds to Art 22 of the GDPR stating that decisions made against and effecting individuals in a significant way should not be made using automated systems alone, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance)' [2016], OJ L119 (GDPR).

Code of Conduct.[47] This Code marked a clear starting point for the Commission's work in placing responsibility for the dissemination of illegal content on VLOPs.[48]

It is worth noting that 'illegal content' is broadly defined in the DSA,[49] and is primarily determined by national legislation, and in some cases, EU law. Recital 12 of the DSA, however, provides general examples of illegal material, such as hate speech, terrorist content and material depicting child sexual abuse. It also clarifies that the online environment should reflect the same legal boundaries as the offline world.

The DSA explicitly incorporates the Code of Conduct when outlining the responsibilities of VLOPs in managing risks related to illegal content uploaded to their platforms. For instance, the Code's 24-hour rule for the removal of suspected hate speech is used as a benchmark. Nonetheless, the DSA states in Recital 87 that certain content may require either more or less time to process, depending on its nature, context, and the facts involved. The exact length of these timeframes is not specified. While this demonstrates an awareness of the need for contextualisation, it offers little concrete guidance for VLOPs on how to achieve it. Recital 87 also notes that '[o]ther appropriate measures could include adapting their content moderation systems and internal processes or adapting their decision-making processes and resources, including the content moderation personnel, their training and local expertise.'[50] This can be understood as a recognition of the importance of well-trained human moderators with knowledge of local contexts to ensure informed decisions.

The 24-hour timeline also appears in the DSA's provisions on trusted flaggers – entities described in Article 22 as having 'particular expertise and competence for the purposes of detecting, identifying and notifying illegal content'.[51] Trusted flaggers are independent from any platform and are expected to operate objectively. Their involvement is seen as a way for platforms to speed up moderation processes, particularly concerning content involving minors or other vulnerable groups. However, it is again stated that certain content may require longer to process,[52] though no detailed explanation is provided. Implicit in this is a need for human oversight or referral to internal moderation systems to allow both contextualisation and timely decision-making.

---

[47] Recitals 87 and 62 DSA. The Code of Conduct was first signed in 2016 by the Commission and Facebook, Microsoft, Google (including YouTube) and Twitter (now X), and has since been joined by other major platforms such as Twitch, Snapchat, LinkedIn and TikTok.

[48] This is done by still clearly expressing the significance of placing individual responsibility on perpetrators and highlighting the importance of national criminal law sanctions. European Commission (n 10); with references to Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law [2008] OJ L 328.

[49] Art 2(h) DSA.

[50] Recital 87 DSA.

[51] Art 22(a) DSA.

[52] Recital 62 and Art 22 DSA.

The DSA also requires VLOPs to implement notice-and-action systems, enabling any user to electronically notify the service provider of potentially illegal content in a clear and structured manner.[53] This mechanism ensures platforms are made aware of content that may warrant moderation. According to the DSA, notice-and-action systems must be sufficiently precise to allow platforms to determine whether the reported content should be restricted or removed. These systems are directly tied to the platform's perceived 'awareness' under the DSA. If a notice is sufficiently clear – such that the illegality of the content is apparent even without a detailed legal analysis – it is deemed that the platform has been made aware.[54] From that moment, the DSA imposes an obligation to act swiftly to remove or restrict access to that content.[55] Moreover, transparency requirements apply: individuals or entities reporting content through notice-and-action systems must be informed of the outcome so that they may respond appropriately.[56]

It is important to understand how these notices are handled in a manner that meets the preferred moderation timeframe, while also maintaining transparency and upholding fundamental rights. For example, TikTok has outlined how it ensures compliance with the DSA. When content is reported as potentially illegal, it is first reviewed against TikTok's internal guidelines and removed if found in violation. If the content does not breach the platform's policies, it is reviewed against local legislation – by either a moderator or legal specialist. Both the reporting user and the content creator are notified of the outcome, and an appeal may be lodged.[57]

On Meta's platforms (Facebook and Instagram), users may also appeal moderation decisions – whether about content that was removed or content that was allowed to

[53] Art 16 DSA.
[54] Recital 53 DSA.
[55] Recital 22 DSA. Other than notice and action systems a provider of an online service can be viewed as having awareness of illegal content after investigations on their own initiative, but never through a general awareness of illegal content being spread using their service. Despite not being the focus in this article, it is worth mentioning that in that sense, the DSA offers a updated wording for the safe harbor regulation on liability for unlawful user-generated content in the e-Commerce directive, Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') [2000], and reaffirms that the exemption from liability when performing voluntary investigations into potential illegal content (Arts 6–7 DSA).
[56] Art 16(5) DSA.
[57] As to be found in TikTok's own information to users on how to report illegal content and how the reports are handled TikTok, 'Reporting Illegal Content' <https://www.tiktok.com/legal/page/global/reporting-illegal-content/en> accessed 3 May, 2025.

remain.[58] As discussed earlier, the DSA requires that such appeal mechanisms be in place.[59]

The right to appeal moderation decisions is designed to strengthen the rights of individual users and to guard against unjustified censorship or unnecessary infringements on freedom of expression. For that reason, the basis of each decision must be made clear, along with a description of the type of moderation system used. In support of transparency, the European Commission has established a DSA Transparency Database, where VLOPs report moderation decisions. One example from TikTok illustrates how such reports may be phrased:

> We do not allow any hateful behavior, hate speech, or promotion of hateful ideologies. This includes content that attacks a person or group because of protected attributes. We proactively enforce our Community Guidelines through a mix of technology and human moderation. We have detected this policy violation using automated measures. We have used automated measures in making this decision.[60]

This content was flagged as violating terms related to hate speech and was subsequently removed. The form specifies that the decision was made using fully automated systems, initiated by the platform.

When content is reported on Facebook, Meta reviews it in accordance with community guidelines, which it states are 'based on feedback from people and the advice of experts in fields like technology, public safety and human rights'.[61] If content is found to be in breach, it is said to be removed.[62] Meta also notes that both automated and human moderation are used to review and identify non-compliant or offensive material. Human moderators are involved when a deeper review is triggered by the AI or machine learning system:

> Sometimes, a piece of content requires further review and our AI sends it to a human review team to take a closer look. In these cases, review teams make the final decision, and our technology learns and improves from each

[58] As to be found in Meta's own information, Meta, 'How to Appeal to the Oversight Board | Transparency Centre' <https://transparency.fb.com/en-gb/oversight/appealing-to-oversight-board> accessed 6 April 2025.
[59] Recital 58 DSA.
[60] European Commission, 'Statement of Reasons Details – DSA Transparency Database' <https://transparency.dsa.ec.europa.eu/statement/36070318306> accessed 6 April 2025.
[61] Facebook, 'Community Standards | Transparency Centre' <https://transparency.fb.com/sv-se/policies/community-standards> accessed 6 April 2025.
[62] Facebook, 'Reporting Abuse | Help Centre' <https://www.facebook.com/help/1753719584844061/?locale=en_GB&helpref=hc_fnav&cms_id=1753719584844061> accessed 6 April 2025.

> decision. Over time – after learning from thousands of human decisions – the technology gets better.[63]

This interaction between automated and human moderation is notable. It implies that AI acts as the first line of moderation, capable of deferring to human moderators when needed. However, as previously noted, automated systems face limitations – particularly in detecting intent, interpreting evolving slang or coded language, and making nuanced judgements. There remains little clarity on how well automated systems perform in such contexts, or how reliably they can identify when human intervention is necessary.

These examples illustrate that platforms – and especially VLOPs – acknowledge the necessity of using both human and automated moderation. This dual approach allows them to manage the volume of content while preserving the contextual analysis needed to protect fundamental rights such as freedom of expression. Notably, this practice is being implemented despite the fact that the DSA does not explicitly require human moderation to be included in content decisions.

2.2.3 Moderating Terrorist Content Online – Demanding Human Contextualisation?

A specific subcategory of illegal content online is the dissemination of material that may be perceived as terrorist propaganda. This issue has been a high priority for the EU in recent years. In 2021, TERREG was adopted and has been applicable since June 2022. The regulation functions as *lex specialis* to the DSA. While the DSA provides limited guidance on when or how automated or human moderation should be applied in the context of terrorist propaganda, TERREG is more explicit:

> Where specific measures involve the use of technical measures, appropriate and effective safeguards, in particular through human oversight and verification, shall be provided to ensure accuracy and to avoid the removal of material that is not terrorist content.[64]

In the recitals it is further stated that platforms must provide safeguards 'including human oversight and verifications, to avoid any unintended or erroneous decision leading to the removal of or disabling of access to content that is not terrorist content'.[65] Additionally, platforms must be able to demonstrate that they have the means to implement sufficient measures to combat terrorist content online – such as through human oversight in content moderation.[66] However, this does not exclude

---

[63] To be found at Facebook, 'How Does Facebook Use Artificial Intelligence to Moderate Content? | Help Centre' <https://www.facebook.com/help/1584908458516247?locale=en_GB&cms_id=1584908458516247> accessed 6 April 2025.
[64] Art 5(3)(d) Regulation (EU) 2021/784; see also Recital 24. This corresponds to Art 22 GDPR.
[65] Recital 23 Regulation (EU) 2021/784.
[66] Recital 24 Regulation (EU) 2021/784.

the use of automated moderation if it is deemed an effective tool for combating terrorist content.[67]

Moderating terrorist propaganda presents numerous challenges. Beyond the emotional toll on moderators exposed to extreme content,[68] the contextual nature of potential terrorist propaganda makes it difficult for any moderation system to assess. The highly politicised environments in which such content often circulates heighten the risk of infringing upon fundamental rights, especially freedom of expression.[69] Moderation systems must be capable of identifying specific acts. TERREG adopts the same definition of terrorism as the EU Terrorism Directive,[70] but focuses specifically on the online environment and on acts that can be carried out online, excluding all physical acts of terrorism.

The scope of what qualifies as terrorism content is nonetheless broad, including material that incites or solicits others to commit physical attacks, kidnappings, or to release dangerous substances, as well as, for example, instructions on how to construct explosives.[71] Some guidance is provided for platforms in assessing whether content should be considered terrorist in nature, including:

> [F]actors such as the nature and wording of statements, the context in which the statements were made and their potential to lead to harmful consequences in respect of the security and safety of persons.[72]

This recognises the need for highly contextualised moderation decisions, but stops short of offering a concrete framework for making such assessments – undoubtedly due to the complexity and discretion such decisions require.

TERREG's complexity is acknowledged further in Recitals 11 and 12. Recital 11 affirms that the need to address the most harmful terrorist propaganda justifies a broad scope of regulation. In contrast, Recital 12 explicitly excludes content intended to raise awareness or combat terrorism, as well as material shared for educational, journalistic, or artistic purposes. Content disseminated for research is also excluded from the scope of terrorist content. Recital 12 additionally clarifies that 'expression

---

[67] Recital 25 Regulation (EU) 2021/784.
[68] Roberts (n 18); Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018) 11.
[69] Eugénie Coche, 'Countering Terrorism Propaganda Online Through TERREG and DSA: A Battlefield or a Breath of Hope for Our Fundamental Human Rights?' in Dário Moura Vicente, Sofia de Vasconcelos Casimiro and Chen Chen (eds), *The Legal Challenges of the Fourth Industrial Revolution* (Springer International Publishing 2023).
[70] Art 2(7) Regulation (EU) 2021/784; see also Recital 11.
[71] Art 2(7) Regulation (EU) 2021/784, to be read with Art 3(2)(p)(a–l) of Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA 2017.
[72] Recital 11 Regulation (EU) 2021/784.

of radical, polemic or controversial views in the public debate on sensitive political questions should not be considered to be terrorist content'.

VLOPs must therefore ensure that their moderation systems can distinguish between, for example, political statements that could be interpreted as inciting violence and those that are merely radical or polemic. Systems must also differentiate between videos depicting terrorist acts intended to raise awareness of atrocities, and those that glorify the actions shown. This balancing act is inherently difficult, and there is currently no case law to guide the interpretation or practical implementation of these requirements. Moreover, users may find it difficult to understand the measures taken to maintain this balance.

As with the DSA, TERREG requires that hosting services – such as VLOPs – clearly articulate in their terms and conditions what users are agreeing to, including commitments to refrain from sharing content that may constitute terrorist propaganda, and whether automated moderation is used to counter such content.[73]

Although TERREG acts as *lex specialis* to the DSA, VLOPs may still be held accountable under the DSA for failing to moderate terrorist content if they are shown to have knowledge of its presence on their platforms.[74] Once VLOPs become aware – by any means – that such content may exist on their service, they are obligated to act. Failure to comply may result in liability under the DSA, including fines of up to six per cent of global turnover.[75] This underscores both the significance of the DSA for platforms and the tensions that arise from the somewhat vague obligations under TERREG and their interplay with the DSA.

2.2.4 Moderating Disinformation – Moderation or Fact-checking?

Disinformation represents another complex societal phenomenon. Unlike terrorist content, the EU has adopted a different approach to regulating disinformation, focusing on self- and co-regulatory measures in combination with general legislative obligations under the DSA. As previously discussed, alongside initiatives to combat hate speech and terrorist propaganda, the EU has simultaneously taken steps to address disinformation. It is not always easy to distinguish between hateful content, terrorist propaganda and disinformation. These categories can overlap, and while the former two may also involve disinformative elements, many other types of content can be classified as disinformation.

The DSA does not offer a formal legal definition of disinformation. Like *illegal content*, its identification is primarily left to national legislation, although some EU-level laws

---

[73] Art 7(1) and Recital 11 Regulation (EU) 2021/784.

[74] Here it is important to note that nothing in the DSA, TERREG or the Terrorism directive subject VLOPs to an obligation to apply general monitoring, detection and moderation of illegal material, including terrorism material, on their platforms. On the contrary, all these regulatory measures specifically clarify that VLOPs do not have that obligation; see Art 8 and Recital 30 DSA; Recital 23 Directive (EU) 2017/541; and Art 5(8) (and Recital 25) Regulation (EU) 2021/784.

[75] Art 74(1) DSA.

may also apply. Several Member States have already enacted legislation to counter or prohibit disinformation.[76]

Despite the absence of a clear legal definition in the DSA, VLOPs are still subject to several obligations related to disinformation. Advertising is specifically addressed in this context, particularly regarding the vulnerability of individuals targeted by ads tailored to their personal interests.[77] Disinformation is also referenced in DSA provisions concerning content moderation. This includes a focus on automated moderation systems or algorithmic tools that promote, demote or remove content based on user interaction. The DSA states that VLOPs should undertake 'corrective measures', such as promoting information from authoritative sources and raising user awareness when disinformation campaigns are identified on their platforms.[78]

In assessing risks, platforms are required to evaluate their moderation systems and available resources, as well as how their services may facilitate the dissemination of disinformative content. This includes content that is *not in itself illegal* but may contribute to systemic risks:

> 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. Such providers should therefore pay particular attention on how their services are used to disseminate or amplify misleading or deceptive content, including disinformation.[79]

While the DSA offers little specificity on how content moderation should be conducted to combat disinformation, it encourages both self- and co-regulatory measures. The Regulation identifies disinformation as one of the societal harms that justifies regulatory focus.[80] Here, the DSA stresses that 'adherence to and compliance with'[81] a code of conduct may be seen as appropriate measures taken by VLOPs. The Code of Practice is specifically referenced in this regard.[82]

---

[76] Like Germany with The Network Enforcement Act 'NetzDG – Gesetz Zur Verbesserung Der Rechtsdurchsetzung in Sozialen Netzwerken' <https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html> accessed 7 March 2025, and France with LOI n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (1) 2018 (2018-1202) stating that judges can authorise removal of political disinformation in relation to election campaigns.

[77] This is addressed in the Recital 69 DSA, and refers to personal information as defined in Art 4(4) GDPR, with reference to Art 9(1).

[78] Recital 88 DSA.

[79] Recital 84 DSA.

[80] This is mentioned alongside actions that manipulate or abuse minors.

[81] Recital 104 DSA.

[82] Recital 106 DSA. Similar initiatives are highlighted in Recital 88. In the European Commission, '2022 Strengthened Code of Practice on Disinformation' (16 June 2022), it is stated that the code is complementary to and aligns with the DSA; see preamble [H].

The Code of Practice builds on the original 2018 version. The 2018 document marked the first time industry actors agreed to tackle disinformation through a self-regulatory framework.[83] The updated 2022 version includes a broader range of signatories and aims to combat disinformation across diverse online environments, in response to calls from the European Commission for stronger action.[84]

As noted, the Code of Practice became a formal code of conduct under the DSA in 2025. It provides guidance on the types of content considered disinformation and outlines recommended actions. A key distinction is made between misinformation and disinformation: both involve the dissemination of false or misleading information, but disinformation is defined by the presence of deliberate intent to deceive. Misinformation, by contrast, lacks this intentional element. Nonetheless, both categories fall within the scope of the Code. It also encompasses both domestic and foreign influence campaigns designed to manipulate public opinion, as well as efforts by foreign state actors to disrupt democratic processes – following encouragement from the Commission for their inclusion.[85] The avoidance of defining disinformation in the DSA was perhaps by design, since the Code of Practice was always intended to be a code of conduct under the DSA.[86]

Signatories to the Code[87] commit to offering users tools to verify the accuracy of information – such as third-party fact-checking services.[88] These tools are intended to help users navigate the vast volume of available information without relying solely on moderation systems or human moderators to determine what should be removed. In fact, direct content moderation is rarely addressed in the Code. When it is, it tends to concern the reliability of technological tools, such as AI used to detect content like deepfakes (see Commitments 15 and 16 of the Code). The Code also promotes the sharing of best practices among signatories on moderation and related activities to more effectively counter mis- and disinformation.[89]

The Code further highlights the importance of ensuring that content flagging systems are secure and not vulnerable to manipulation – whether by human users or

---

[83] European Commission, '2018 Code of Practice on Disinformation | Shaping Europe's Digital Future' (16 June 2022) <https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation> accessed 7 March 2025.
[84] European Commission, '2022 Strengthened Code of Practice on Disinformation | Shaping Europe's Digital Future' (16 June 2022) <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation> accessed 7 March 2025.
[85] Ibid 1(a) with footnotes; European Commission, 'Statement of Reasons Details − DSA Transparency Database' <https://transparency.dsa.ec.europa.eu/statement/36070318306> accessed 6 April 2025.
[86] Code of Practice (n 84) 1(i).
[87] European Commission, 'Signatories of the 2022 Strengthened Code of Practice on Disinformation | Shaping Europe's Digital Future' (16 June 2022) <https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation> accessed 7 March 2025.
[88] Code of Practice (n 84), Commitments 21 and 22.
[89] Ibid Commitments 15 (measure 15(2)) and 16 (measure 16(2)).

automated systems – that could lead to the undue silencing of individuals. Again, this emphasis is not directly tied to traditional content moderation. Much of the responsibility for managing disinformation is envisioned as resting with users themselves, who are empowered through various platform tools.

A clear example of this is the Code's emphasis on 'empowering' three key groups: users; fact-checkers; and the research community. This approach aims to foster safe and transparent platforms where users can report questionable content, access assistance in evaluating information, and – consistent with the other regulatory instruments discussed in this article – appeal moderation decisions made against their content or accounts.[90]

From this, it becomes apparent that the question of how content moderation is to be conducted is only moderately addressed. This does not suggest that the seriousness or complexity of disinformation is underestimated; rather, the Code focuses on alternative strategies, including fact-checking tools and appeals mechanisms. The difficulty of defining disinformation with precision underscores the need for support from third-party fact-checkers, both to help address systemic threats and to create a safer environment for users.

The DSA actively promotes voluntary codes of conduct and similar frameworks to counter disinformation, while retaining binding obligations that hold VLOPs accountable for non-compliance. As noted, if a VLOP adheres to a code of conduct, this may be considered indication that it has taken appropriate steps – including moderation – toward limiting the spread of disinformation on its platform.[91]

3. Drawing the Map of Emerging Content Moderation Requirements

The issue of combating illegal or harmful content online has never received more attention in the EU than it does today, with a range of regulatory instruments introduced that complement – and at times overlap with – one another. It is also important to remember that the development and interpretation of these instruments will be closely intertwined with the evolution of case law from both the CJEU and ECtHR on content moderation and freedom of expression. As discussed in section 2.1, existing case law already demonstrates the complexity of applying a contextualised understanding to content moderation in order to avoid undue infringements on fundamental rights, while also placing responsibilities on VLOPs to

[90] Ibid see Commitments 18, 21, 24, 26, 31 and 32. Other examples are demanding that signatories provide independent fact-checkers with information that help them conduct efficient fact-checking, and the research community with data that enables sufficient research on disinformation. Providing an application programming interface (API) is one example.

[91] M.R. Leiser, 'Reimagining Digital Governance: The EU's Digital Service Act and the Fight Against Disinformation' (24 April 2023) 8–9 <https://papers.ssrn.com/abstract=4427493> accessed 7 March 2025.

automatically delete certain types of content. The DSA itself must be seen as part of a highly complex regulatory framework.

Although VLOPs are under organisational obligations to provide safe and transparent spaces for users and to prevent the amplification of societal risks on their platforms, the DSA refrains from prescribing how content moderation processes should be structured. Nonetheless, the DSA effectively demands advanced moderation systems. To be efficient and as accurate as possible, such systems must rely on both human moderators and automated tools. The structure of moderation processes is left largely to the discretion of VLOPs, so long as they implement moderation and provide users with safeguards, including mechanisms to appeal decisions. As previously mentioned, the DSA includes specific provisions requiring access to human review – but only when users initiate a challenge to a moderation decision. Similarly, TERREG outlines the importance of human oversight as a safeguard to ensure that fundamental rights are not unduly infringed. However, this does not prohibit automated moderation, provided that appropriate safeguards are in place. This offers platforms substantial flexibility in designing their moderation systems. In any case, requiring human review for most moderation decisions would be practically unfeasible due to the volume of content being published.[92] The primary emphasis, therefore, is on transparency regarding how decisions are made.

The responsibility imposed on VLOPs is thus shared with their users, who are indirectly tasked with contributing to the moderation ecosystem – for instance, through appeals or content reporting mechanisms. In the context of TERREG, human oversight is especially important given the highly political and contextual nature of suspected terrorist content and the risks it poses to society and individuals alike. Both removal and non-removal of such content can have serious implications for safety and rights. VLOPs must therefore be able to distinguish between malicious content inciting terrorist acts and content intended to raise awareness of terrorism, as well as between radical but lawful speech and unlawful incitement. These assessments – particularly the intention behind shared content – are complex, even for human moderators, and significantly more so for automated systems.

The demanding moderation environment also underlines the importance of using trusted flaggers and independent experts, as highlighted in the DSA, and fact-checkers in relation to disinformation in the Code of Practice. Involving actors beyond traditional moderation systems – human or otherwise – is essential to guide and inform these systems. The particularly sensitive categories of terrorist propaganda and disinformation exemplify the difficulty of moderating such content while respecting users' fundamental rights to access and share information, and the need

[92] However, this means that, despite the demands for transparency, the actual moderation, often automated, is often time opaque for the users. They can access information on what grounds the decision to use a moderating action was made (for example, hate speech or hateful behaviour in breach with TikTok's policy as previously cited), but when initiated on platforms' own voluntary initiative, the reason for that investigation can remain unclear.

to protect them from harm. Automated systems are necessary for efficiency and for reducing the exposure of human moderators to harmful material, but platforms must be granted some leeway in determining how to meet the goals set out in the DSA and related regulatory instruments.

This article has sought to contribute to understanding how recent EU instruments regulate VLOPs' responsibilities for effective content moderation, with specific focus on balancing speed and contextualisation. The findings show that this interrelationship indirectly influences the regulation. As discussed, the DSA explicitly requires human moderation only when users appeal decisions – meaning that, in most cases, moderation will occur entirely through automated systems, even for highly sensitive content. Human oversight is not generally presented as an essential safeguard during the initial decision-making process. Nonetheless, indirect requirements for human moderation exist, as platforms often refer complex cases to human reviewers. Importantly, the ability to recognise when a case requires human intervention already presupposes a degree of (human) contextual and communicative understanding.

There are currently no binding rules or guidance specifying content types or circumstances that always require human review. According to the DSA Transparency Database approximately 49% of moderation decisions are made entirely through automated systems.[93] Here, further interpretation by the CJEU may help clarify the importance of human moderation in fulfilling the DSA's objectives. In particular, the type of contextualisation required in politically sensitive contexts – such as armed conflicts or terrorism – may be difficult or impossible to achieve through automated systems alone. Even if VLOPs do in practice refer such content to human reviewers, this could be formalised through regulation.

It is also important to note that many of the regulatory instruments relevant to content moderation are voluntary and rely on self- or co-regulation. For instance, the Code of Practice on Disinformation is voluntary, and signatories may withdraw at any time. The VLOP X chose to do so in May 2023. However, only a few months after the DSA came into force for VLOPs, the first formal proceedings were initiated against the same platform in December 2023. This followed earlier requests from the European Commission to X regarding the suspected dissemination of illegal and disinformative content – particularly terrorist material, violent content and hate speech – in the aftermath of Hamas's terrorist attack on Israel.[94] The Commission opened a formal investigation focusing on several issues, including obligations under Articles 34 and

[93] European Commission, 'DSA Transparency Database' <https://transparency.dsa.ec.europa.eu > accessed 8 April 2025.
[94] European Commission, 'The Commission Sends Request for Information to X under DSA' <https://ec.europa.eu/commission/presscorner/detail/en/ip_23_4953> accessed 7 March 2025.

35 DSA relating to risk mitigation and the adequacy of content moderation resources.[95]

This illustrates the strength of the DSA's enforcement framework, even though it remains vague on the specific types of moderation required. In contrast, voluntary regulatory instruments such as codes of conduct may offer clearer guidance on tools and practices for moderating specific types of content. The Commission's action against X may represent an early step in clarifying responsibilities relating to risk assessment and mitigation, and the adequacy of moderation capacity.[96]

Such guidance is urgently needed. Overall, there is little direction given to VLOPs (or to other platforms) on how to conduct content moderation in a way that balances the protection of individuals and societies from harmful content with the right to receive and share information under freedom of expression. While the legal obligations placed on VLOPs recognise their vital role as internet gatekeepers, they are generally phrased in open-ended terms that leave significant discretion. The DSA encourages transparency and the development of voluntary best practices, such as those reflected in the Code of Conduct, which may, in turn, inform future iterations of the DSA.

At the same time, the use of self-regulatory mechanisms and the enforcement of community guidelines may also serve as a strategic workaround for platforms, enabling them to moderate content more freely and maintain alignment with user expectations. This may serve the dual purpose of protecting users while preserving the platform's character and preventing over-moderation.

Finally, one possible reason for the EU's reluctance to define specific categories of removable content could be to avoid constitutional challenges about whether the correct balance has been struck between competing rights. Instead, the EU has adopted a structural model that requires platforms to implement transparent policies and moderation systems while leaving room for national-level interpretation and enforcement. This also allows public and private actors in Member States to hold VLOPs accountable where moderation is found to be insufficiently effective.

Regardless of the systems employed, these regulatory instruments recognise the importance of preserving content that is hidden due to suspected illegality or guideline breaches. This facilitates the reinstatement of wrongfully removed content and preserves potential evidence for criminal investigations. Such records may also help shape future regulatory developments and inform platforms' own self-regulatory approaches.

---

[95] European Commission, 'Commission Opens Formal Proceedings against X under the DSA' <https://ec.europa.eu/commission/presscorner/detail/en/IP_23_6709> accessed 7 March 2025. Some preliminary findings were presented in 2024, European Commission 'Commission sends preliminary findings to X for breach of DSA' <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761> accessed 8 April 2025.

[96] Ibid.

In conclusion, drawing a map of emerging content moderation requirements reveals many areas that remain uncharted and uncertain. For the benefit of both users and VLOPs, further research is needed into the evolving responsibilities that define this regulatory shift in online content governance. While the DSA and related frameworks only apply within the European market, their influence is likely to extend beyond the EU.[97] The principles underpinning the DSA – especially its novel approach to defining VLOP responsibility – signal a shift in platform regulation and user protection, underscoring the continued need to monitor and analyse the implications of the DSA for online content moderation.

[97] Thales Martini Bueno and Renan Gadoni Canaan, 'The Brussels Effect in Brazil: Analysing the Impact of the EU Digital Services Act on the Discussion Surrounding the Fake News Bill' (2024) *Telecommunications Policy* 102757; Martin Husovec, 'Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation of Global Internet Rules' (2023) 38 *Berkeley Technology Law Journal* (forthcoming) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4598426> accessed 2 May 2025.