Computer Law & Security Review 59 (2025) 106219

Contents lists available at ScienceDirect

Computer Law & Security Review: The International Journal of Technology Law and Practice

journal homepage: www.elsevier.com/locate/clsr

The digital prior restraint: Applying human rights safeguards to upload filters in the EU

Emmanuel Vargas Penagos

School of Behavioural, Social and Legal Sciences, Örebro University, Örebro, Sweden

ARTICLE INFO

Keywords:
Upload filters
freedom of expression
Social media
Digital services act
TERREG
Copyright directive
Prior restraints
Very large online platforms
VLOPs

ABSTRACT

This article examines the human rights standards relevant to the use of upload filters for content moderation within EU secondary legislation. Upload filters, which automatically screen user-generated content before publication, are a type of prior restraint, which raises critical concerns on freedom of expression. EU secondary legislation establishes rules for both mandatory and voluntary use of these technologies, which must be read in light of human rights protections. This article analyses the characteristics of both mandatory and voluntary upload filters as prior restraints, the relevant EU legal provisions governing their use, and the safeguards required to prevent disproportionate restrictions on speech. Additionally, it explores the procedural and institutional safeguards under EU law, viewed through the lens of the CJEU and ECtHR case law on prior restraints and the rights to a fair trial and to an effective remedy.

1. Introduction

On 30 June 1971, three U.S. Supreme Court judges criticized the rushed timeline (less than a week) that they had been given to decide whether the U.S. government could block *The New York Times* and *The Washington Post* from publishing a damning report on the Vietnam War. They warned that such 'difficult questions of fact, of law, and of judgment' required more time.[1] Their concern about the complexity and urgency of ruling on a single press publication illustrates the stakes in a world where such rapid, high-stakes decisions are now made by 'upload filters' deployed by social media platforms at the scale of millions. Almost half a century later, some of these companies boast of their use of such technologies to block content 'before a single human user has ever been able to access it'.[2] Beyond this anecdote, a parallel debate has unfolded in the EU legislature over laws seen as imposing, enabling or not addressing the concerns of such preemptive censorship. In this context, some have argued that 'we all know they are already used by big platforms' and that rejecting legislative reform 'will not take away upload filters'.[3] Others have expressed disappointment at having 'failed to protect legal content, including media content, from being over-blocked by error-prone upload filters or arbitrarily set platform rules'.[4] Based on existing EU legislation and on the applicable standards developed by both the European Court of Human Rights (ECtHR) and the Court of Justice of the European Union (CJEU), this paper reflects on the rules and principles that should govern the use of such technologies by social media companies.

Upload filters, namely automated tools deployed by internet intermediaries (principally social media platforms), to preemptively

E-mail address: emmanuel.vargas-penagos@oru.se.

[1] *New York Times Co v United States* 403 U.S. 713 (1971).

[2] US Congress, *Fostering a Healthier Internet to Protect Consumers: Joint Hearing before the Subcommittee on Communications and Technology and the Subcommittee on Consumer Protection and Commerce of the Committee on Energy and Commerce, House of Representatives, One Hundred Sixteenth Congress, First Session, October 16, 2019* <http://catalog.gpo.gov/F/?func=direct&doc_number=001167713&format=999> accessed 28 February 2025.

[3] See the words given by Andrus Ansip, Vice-President of the Commission at 'Verbatim Report of Proceedings - Copyright in the Digital Single Market (Debate) - Tuesday, 26 March 2019' <https://www.europarl.europa.eu/doceo/document/CRE-8-2019-03-26-ITM-002_EN.html> accessed 28 February 2025.

[4] See the words given by MEP Patrick Breyer at: 'Verbatim Report of Proceedings - Digital Services Act - Digital Markets Act (Debate) - Monday, 4 July 2022' <https://www.europarl.europa.eu/doceo/document/CRE-9-2022-07-04-ITM-015_EN.html> accessed 28 February 2025.

https://doi.org/10.1016/j.clsr.2025.106219

Available online 8 October 2025

examine user-generated content to find material that could be deemed to be unlawful or against their terms and conditions, have been encouraged or demanded by different stakeholders. These include such diverse voices as governments combating terrorism,[5] and copyright holders seeking to prevent the unauthorised distribution of music, films and series.[6] These stances have not escaped resistance from civil society,[7] academia,[8] internet pioneers[9] and human rights bodies, including the United Nations Special Rapporteur on freedom of opinion and expression (the UN Foe Rapporteur).[10] Their argument is that upload-filters amount to a 'prior restraint' and therefore pose a threat to freedom of expression. At the European level, prior restraints have traditionally been considered as restrictions that pose an inherent danger to freedom of expression. Despite not being prohibited outright, they are called to be examined with the most careful scrutiny[11] under a strict legal framework.[12]

Despite this resistance to upload filters, the stance taken at the EU level, as Barral Martinez notes, places their use in content moderation as 'a de facto must for online intermediaries to escape liability, especially in cases where short removal time is required.'[13] This may be reinforced by the recent position taken by the ECtHR that 'a minimum degree of subsequent moderation or automatic filtering would be desirable in order to identify clearly unlawful comments as quickly as possible and to ensure their deletion within a reasonable time, even where there has been no notification by an injured part'.[14] Such a view is likely to influence the implementation of EU legislation, particularly because Article 11 of the EU Charter[15] and Article 10 ECHR,[16] both enshrining freedom of expression, have the same meaning and scope and have corresponding interpretations.[17] This must be read in line with the CJEU's case law, which, embracing previous ECtHR rulings, has reasoned that upload filters require 'a particularly tight legal framework' to avoid encroachments on freedom of expression.[18] Moreover, this should also be considered in light of the ECtHR's reasoning that, in the context of prior restraints, it is necessary to make a 'close examination of the procedural safeguards embedded in the system' to prevent arbitrary restrictions on free expression.[19] In that sense, it will also be relevant to examine the case law from both courts in relation to the right to a fair trial and the right to an effective remedy (safeguarded by Articles 6 and 13 ECHR and Article 47 of the Charter, respectively), which, at the same time, have also been considered to be corresponding provisions by the CJEU.[20] In contrast, platforms may sometimes use filters to ensure a better experience to their users by, for instance, preventing the dissemination of spam.

The EU legal framework has several pieces of sectoral content moderation-related legislation, such as the Audiovisual Media Services Directive (AVMSD),[21] the Copyright Directive,[22] the Regulation on addressing the dissemination of terrorist content online (TERREG),[23] the Platform to Business Regulation,[24] the European Media Freedom Act[25] and the Directive on combating violence against women and domestic violence (VAW Directive),[26] and so on. Beyond this, the Digital Services Act (DSA)[27] seeks to provide an overarching general framework with rules addressed to internet intermediaries, such as social media, for the effective protection of fundamental rights. These rules have a significant impact on whether and how upload filters are applied and, given the equivalent meaning and scope between article 11 of the EU Charter and Article 10 of the ECHR, their reading requires having regard to the relevant human rights standards developed at the ECtHR.

Against this background, the purpose of this article is to provide an overarching view on how the procedural safeguards developed by the ECtHR apply in the context of 'upload filters' within the scope of relevant EU secondary legislation. Therefore, this article addresses the following question: *How can human rights procedural safeguards against interferences on freedom of expression be applied to adequately protect*

[5] Amélie Pia Heldt, 'Upload-Filters: Bypassing Classical Concepts of Censorship?' (2019) 10 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 56.

[6] Annemarie Bridy, 'The Price of Closing the Value Gap: How the Music Industry Hacked EU Copyright Reform' (2020) 22 Vanderbilt Journal of Entertainment & Technology Law 323.

[7] Civil Liberties Union for Europe, 'Article 13 Open Letter – Monitoring and Filtering of Internet Content Is Unacceptable' (*Liberties.eu*, 16 October 2017) <https://www.liberties.eu/en/stories/delete-article-thirteen-open-letter/13131> accessed 31 January 2025.

[8] Martin Kretschmer, 'The Copyright Directive: Misinformation and Independent Enquiry – CREATe' (24 March 2019) <https://www.create.ac.uk/blog/2018/06/29/the-copyright-directive-misinformation-and-independent-enquiry/> accessed 31 January 2025.

[9] Danny O'Brien and Jeremy Malcolm, '70+ Internet Luminaries Ring the Alarm on EU Copyright Filtering Proposal' (*Electronic Frontier Foundation*, 12 June 2018) <https://www.eff.org/deeplinks/2018/06/internet-luminaries-ring-alarm-eu-copyright-filtering-proposal> accessed 31 January 2025.

[10] David Kaye, 'EU Must Align Copyright Reform with International Human Rights Standards, Says Expert' (*OHCHR*, 11 March 2019) <https://www.ohchr.org/en/news/2019/03/eu-must-align-copyright-reform-international-human-rights-standards-says-expert> accessed 31 January 2025.

[11] *Observer and Guardian v the United Kingdom* App no 13585/88 (ECtHR, 26 November 1991).

[12] Case C-401/19 *Republic of Poland v European Parliament and Council of the European Union* EU:C:2022:297.

[13] María Barral Martínez, 'Platform Regulation, Content Moderation, and AI-Based Filtering Tools: Some Reflections from the European Union' (2023) 14 JIPITEC <http://www.jipitec.eu/issues/jipitec-14-1-2023/5716>.

[14] *Sanchez v France* App no 45581/15 (ECtHR, 15 May 2023), para 190; See similarly *Zöchling v Austria* App no 4222/18 (ECtHR, 5 September 2023), para 13.

[15] Charter of Fundamental Rights of the European Union [2012] OJ C 326.

[16] Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR).

[17] *Republic of Poland v European Parliament and Council of the European Union* (n 12), para 44.

[18] ibid, para 68.

[19] *Cumhuriyet Vakfi and Others v Turkey* App no 28255/07 (ECtHR, 8 October 2013), para 61.

[20] Case C-487/19 *W.Ż.* EU:C:2021:798, para 123.

[21] Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities [2018] OJ L 303, 28.11.2018, pp. 69–92 (Audiovisual Media Services Directive).

[22] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L 130, 17.5.2019, pp. 92–125 (Copyright Directive).

[23] 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 L 172, 17.5.2021, pp. 79–109 (TERREG).

[24] Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services [2019]OJ L 186, 11.7.2019, pp. 57–79 (Platform to Business Regulation).

[25] Regulation (EU) 2024/1083 of the European Parliament and of the Council of 11 April 2024 establishing a common framework for media services in the internal market and amending Directive 2010/13/EU [2024] OJ L, 2024/1083, 17.4.2024 (European Media Freedom Act).

[26] Directive (EU) 2024/1385 of the European Parliament and of the Council of 14 May 2024 on combating violence against women and domestic violence [2024]OJ L, 2024/1385, 24.5.2024 (VAW Directive).

[27] 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 [2022]OJ L 277, 27.10.2022, pp. 1–102 (DSA).

freedom of expression in the use of upload filters for content moderation under EU law?

To better answer this question, the article will also consider these sub-questions: *i) What are upload filters and what is their impact on freedom of expression?; ii) How does EU secondary legislation address the use of upload filters for content moderation?; iii) What safeguards do the ECtHR and the CJEU provide against prior restraints and how can they be incorporated into the use of upload filters?*

Scholarship on social media companies' use of upload filters in Europe has largely focused on copyright, given that most litigation reaching the CJEU has concerned this issue and that it was central to the heated debate over the adoption of the Copyright Directive.[28] Within this context, research has examined the human rights implications of upload filters, including their classification as a prior restraint. Studies have also explored their use in counterterrorism efforts.[29] Additionally, upload filters are frequently discussed in broader debates on content moderation, where prior-restraint concerns occasionally arise.[30] In that context, while less abundant, research has also addressed upload filters under the Digital Services Act as a cross-cutting regulatory framework for content moderation.[31] Furthermore, research has tended to focus on the mandatory use of upload filters by means of court injunctions, particularly because there is extensive case law on the subject matter,[32] but less focus is given to the voluntary use of those tools by platforms.

While the classification of upload filters as prior restraints is already well-established in the existing literature, this article contributes to these discussions by providing an in-depth analysis of the prior-restraint nature of upload filters and their intersection with human rights safeguards under EU law, highlighting the protections that should apply if and when such filters are implemented. In that sense, this article is not limited to pointing at upload filters as a form of prior restraint but, instead, it analyses the characteristics of prior restraints in case law and literature and explains how they are translated in the context of upload filters.

In addition to this, this article is not limited to the analysis of mandatory upload-filtering as it also examines their use as voluntary measures for content moderation. This examination is particularly relevant given the growing perception of online platforms as de facto or delegated adjudicators of speech[33] and, notably, due to the DSA obligation for platforms to act 'in a diligent, objective and proportionate manner with due regard to the rights and legitimate interests of all parties involved' when restricting speech in their services.[34] It should be noted that the approach within the EU legislature for protecting users from shortcomings in upload filtering focuses on transparency and remedial safeguards. The former are not the focus of this article, but can be summarized as a set of requirements for platforms to disclose information on the automation of content moderation decision-making in their terms and conditions, ,[35] annual reports,[36] notifications of decisions made under notice and action mechanisms,[37] and statements of reasons for restrictions on users based on findings that their content is illegal or violates the platform's terms and conditions.[38] Furthermore, given that the EU's approach to content moderation prioritizes 'procedure before substance,'[39] and the CJEU and ECtHR case law emphasize the need to address prior restraints through a strict legal framework with strong procedural safeguards, this article examines the current remedial framework from a rights-based perspective. In that sense, this article offers a detailed analysis of the institutional and procedural safeguards in place and interprets them in light of CJEU and ECtHR case law on the rights to a fair trial, an effective remedy, and freedom of expression.

It should also be highlighted that this article does not provide a detailed comparison between U.S. and EU law, but refers to relevant aspects of the U.S. framework when useful for analyzing the EU context. This is due to the historical relevance of the U.S. approach on prior restraints and on voluntary measures for content moderation at European level.

This article is structured in three parts. First, it discusses what upload filters are and how they fit within the concept of prior restraint by looking into relevant CJEU and ECtHR case law and literature on the issue, and delineating the impacts on freedom of expression that such considerations have. Second, it explains the way in which mandatory and voluntary upload filters are approached under the applicable EU secondary law and case law by the CJEU. Third, it examines the institutional and procedural safeguards given by the current EU legal framework from a human rights perspective. For this last point, each safeguard is first introduced by explaining its place in the EU legal framework and afterwards provides a reading on the basis of CJEU and ECtHR case law.

[28] Felipe Romero Moreno, '"Upload Filters" and Human Rights: Implementing Article 17 of the Directive on Copyright in the Digital Single Market' (2020) 34 International Review of Law, Computers & Technology 153; Bridy (n 6); Felipe Romero-Moreno, '"Notice and Staydown" and Social Media: Amending Article 13 of the Proposed Directive on Copyright' (2019) 33 International Review of Law, Computers & Technology 187; Christina Angelopoulos and João Pedro Quintais, 'Fixing Copyright Reform: A Better Solution to Online Infringement' (2019) 10 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 147; Giancarlo Frosio, 'To Filter or Not to Filter? That Is the Question in EU Copyright Reform' [2017] SocArXiv <https://ideas.repec.org//p/osf/socarx/67n5w_v1.html> accessed 14 February 2025.

[29] 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*, vol 57 (Springer International Publishing 2023) <https://link.springer.com/10.1007/978-3-031-40516-7_16> accessed 13 August 2024.

[30] Emma J Llansó, 'No Amount of "AI" in Content Moderation Will Solve Filtering's Prior-Restraint Problem' (2020) 7 Big Data & Society 205395172092068; Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945; Niva Elkin-Koren, 'Contesting Algorithms: Restoring the Public Interest in Content Filtering by Artificial Intelligence' (2020) 7 Big Data & Society 2053951720932296; Heldt (n 5); Dawn Nunziato, 'How (Not) to Censor: Procedural First Amendment Values and Internet Censorship Worldwide' [2011] GW Law Faculty Publications & Other Works <https://scholarship.law.gwu.edu/faculty_publications/1286>.

[31] Barral Martínez (n 13); Giancarlo Frosio and Christophe Geiger, 'Taking Fundamental Rights Seriously in the Digital Services Act's Platform Liability Regime' (2023) 29 European Law Journal 31.

[32] Giancarlo F Frosio, 'The Death of "No Monitoring Obligations": A Story of Untameable Monsters' (2017) 8 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 199; Christina Angelopoulos and Martin Senftleben, 'An Endless Odyssey? Content Moderation Without General Content Monitoring Obligations' [2021] SSRN Electronic Journal <https://www.ssrn.com/abstract=3871916> accessed 14 February 2025.

[33] Senftleben, Martin; Quintais, João Pedro; Meiring, Arlette;, 'How the European Union Outsources the Task of Human Rights Protection to Platforms and Users: The Case of User-Generated Content Monetization' (2023) 38 Berkeley Technology Law Journal 933; João Pedro Quintais, Giovanni De Gregorio and João C Magalhães, 'How Platforms Govern Users' Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications' (2023) 48 Computer Law & Security Review 105792; Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136 Harvard Law Review 526.

[34] DSA, art 14(4).

[35] TERREG, art 7(1 ; DSA, art 14(1).

[36] TERREG, art 7(3)(b); DSA, art 15(1).

[37] DSA, art 16(6).

[38] DSA, art 17(3)(c).

[39] Pietro Ortolani, 'If You Build It, They Will Come The DSA "Procedure Before Substance" Approach' in Joris van Hoboken and others (eds), *Putting the DSA into practice* (Verfassungsbooks 2023) <https://doi.org/10.17176/20230208-093135-0>.

Table 1
Non-exhaustive list of upload filtering tools.[47]

Type of technology	Description
Metadata screening	This can be the detection of information constituting the basic description of the file, such as its title, length, size, author, among many others.[1]
Blacklists	Involves a predefined list of banned words, phrases or specific documents that are flagged when uploaded.[2]
Hash databases	Involves creating a database of 'hashes', which are alphanumeric values assigned to specific data, such as content fragments (parts of images, text, video, etc.).[3] When the exact content that was previously uploaded into the database is re-uploaded, it will match the hash in the database and therefore be flagged.[4] However, minor changes to the content, such as altering a pixel of an image, may work to circumvent the detection tool.[5]
Watermarking	Entails introducing specific data, such as barcodes, QRs or stamps to specific files, such as copyrighted works (e.g. films), which allows its detection when uploaded.[6]
Fingerprinting	Used to identify distinctive features within content, such as audio patterns, visual elements, or textual structure, that make it recognizable.[7] Unlike hashing, it allows detection even when the content has been slightly modified. Fingerprinting can be applied to various media types, and aims to either uniquely identify content or match altered versions to the original. Its primary goal is to prevent users from bypassing detection systems through minor edits.[8]
Natural Language Processing (NLP)	Entails using technologies with capacity of conducting a syntactic or semantic analysis of text allowing detection beyond just matching specific words or phrases.[9]
Crowd-intelligence	These techniques can turn useful to moderate content in real time, such as in the case of streaming. This can include examining the comments from viewers to predict the likelihood of infringement from a streaming (e.g. detecting 'clues' relevant to what is meant to be detected, such as references to names of athletes during sports streaming).[10]

[1] Evan Engstrom and Nick Feamster, 'The Limits of Filtering' (Engine 2017) ⟨https://www.engine.is/the-limits-of-filtering/⟩ accessed 3 March 2025.
[2] Policy Department for Citizens' Rights and Constitutional Affairs and Directorate-General for Internal Policies (n 40).
[3] Codecademy Team, 'Hashing: What Is It and How Is It Used?' (*Codecademy Blog*, 28 April 2023) ⟨https://www.codecademy.com/resources/blog/what-is-hashing/⟩ accessed 3 March 2025.
[4] GIFCT, 'Insight: Advances in Hashing for Counterterrorism' (*GIFCT*, 29 March 2023) ⟨https://gifct.org/2023/03/29/advances-in-hashing-for-counterterrorism/⟩ accessed 3 March 2025.
[5] Romero Moreno (n 28); Policy Department for Citizens' Rights and Constitutional Affairs and Directorate-General for Internal Policies (n 40).
[6] Romero Moreno (n 28).
[7] Engstrom and Feamster (n 48); Romero Moreno (n 28); Policy Department for Citizens' Rights and Constitutional Affairs and Directorate-General for Internal Policies (n 40).
[8] Romero Moreno (n 28).
[9] Policy Department for Citizens' Rights and Constitutional Affairs and Directorate-General for Internal Policies (n 40).
[10] Daniel Yue Zhang and others, 'Crowdsourcing-Based Copyright Infringement Detection in Live Video Streams', *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE 2018) ⟨https://ieeexplore.ieee.org/document/8508523/⟩ accessed 14 February 2025.

[47] Author's own work.

2. Upload filters in social media and their impact on freedom of expression

2.1. What are upload filters?

Upload filtering involves automatically and proactively pre-screening user-generated content before publication. Its purpose is to either allow, block, edit, comment on, prioritize, or summarize such content.[40] Other actions that can be the outcome of this pre-screening can include labelling of content or allowing copyright holders to 'monetize' the content unauthorizedly uploaded by third parties.[41] Different types of automated filters conduct this type of *ex ante* moderation by applying predefined rules and leveraging their control over the infrastructure that enables online content publication.[42] While they may help prevent the publication and spread of unwanted content, such as illegal or harmful material, as well as content that may impact the user experience—like spam—they can also introduce delays in publication, impacting both publishers and the public.[43]

While *ex ante* moderation can also be done by human moderators, the term 'upload filtering' refers to the use of automated tools for that same task.[44] The grounds for platforms' decisions to use automated tools for this pre-screening tasks emerge from the large volume of content published on their services, as well as the economic and human costs of hiring teams of human moderators,[45] but can be further reinforced by legal requirements to act promptly within very short timelines.[46]

Social media platforms use a wide range of technical solutions for their moderation processes. In that vein, no single technology inherently qualifies as an upload filter. What defines an upload filter is not the

[40] Policy Department for Citizens' Rights and Constitutional Affairs and Directorate-General for Internal Policies, 'The Impact of Algorithms for Online Content Filtering or Moderation - "Upload Filters"' (2020) PE 657.101.
[41] Senftleben, Martin; Quintais, João Pedro; Meiring, Arlette; (n 33).
[42] James Grimmelmann, 'The Virtues of Moderation' (2015) 17 Yale Journal of Law and Technology 42.
[43] ibid.
[44] ibid.
[45] Vaishali U Gongane, Mousami V Munot and Alwin D Anuse, 'Detection and Moderation of Detrimental Content on Social Media Platforms: Current Status and Future Directions' (2022) 12 Social Network Analysis and Mining 129; Emmanuel Vargas Penagos, 'ChatGPT, Can You Solve the Content Moderation Dilemma?' (2024) 32 International Journal of Law and Information Technology eaae028.
[46] Barral Martínez (n 13).

technology used, but the *point* at which it is used. Since these technologies can also be used *after* publication, the crucial criterion for classifying them as 'upload filters' is that they are applied *before* publication takes place. A non-exhaustive list of the technologies applied can be found in Table 1 below.

In practice, upload filters tend to be used for the detection of spam, phishing, scams, and other types of content that may hinder the user experience.[48] Besides that, they are also used in copyright enforcement, particularly by the implementation of filters that allow the detection of content from the film, TV and music industries, as well as the live streaming of sports. The two most salient technologies in this field are Content ID, developed by YouTube and Audible Magic's Automated Content Recognition, both working with fingerprints of copyrighted content.[49] Furthermore, upload filters are also used very frequently in counterterrorism activities, as well as in countering child sexual exploitation. For the former, the biggest effort is a multiplatform database: The Global Internet Forum to Counter Terrorism (GIFCT), which runs a Shared Industry Hash Database.[50] For the latter, companies use hashing technologies, such as PhotoDNA, which checks uploaded content against a hashing database run by the National Center for Missing and Exploited Children, and Google's CSAI Match and Content Safety API,[51] which are described as being able to detect previously uploaded and previously unseen content, respectively.[52]

Examples of these uses can be seen in some of the transparency reports published by platforms classified as Very Large Online Platforms (VLOPs) under the DSA from 2023 onwards. For instance, X's report of October 2024 says it scans content uploaded to its platform to detect hashes of child sexual abuse images ;[53] TikTok's report covering July to December 2024 says uploads are first reviewed by their 'automated moderation technology, which aims to identify content that violates our Policies before it is viewed or shared by other people on the platform or reported to us' ;[54] Meta's October 2024 reports for Instagram and Facebook do not make reference to upload filtering, but as explained in Section III,[55] they disclose their use for detecting terrorist content in a report mandated by TERREG. Furthermore, Google's DSA report of February 2025 states that they use Content ID to scan for copyrighted content and hashes for detecting child sexual abuse images and terrorist content.[56]

2.2. A brief introduction to prior restraints

The use of upload filters, regardless of the technology or outcome, constitutes an interference with freedom of expression. Indeed, the ECtHR defines an 'interference' as any 'formality, condition, restriction, or penalty' on this right,[57] extending to actions by private parties with or without state control.[58] Similarly, the CJEU holds that platforms conducting 'prior review and prior filtering' through automated tools are liable to restrict online content dissemination, constituting a limitation on Article 11 of the Charter.[59] That said, the fact that interferences to freedom of expression may emerge from private relationships does not mean that the ECHR has a 'horizontal' effect. Instead, it merely entails a State positive obligation to protect convention rights, including freedom of expression, from interferences stemming from private action.[60]

In that vein, upload filters can be seen as a form of prior restraint. This type of interference with freedom of expression has taken various forms throughout history. It has included systems of licensing or permission, such as the book licensing laws enacted after the invention of printing in the 15th century,[61] as well as systems for licensing newspapers, and film licensing boards and requirements for administrative authorisation for distributing pamphlets or organising public assemblies.[62] It has also included different types of injunctions from judicial or administrative bodies to prevent or suppress publications.[63] Licensing or permission systems would usually involve an administrative review where an official decides whether to approve or reject a publication or other type of communicative act, often without a hearing or opposing arguments. The official may also be held responsible if the approved content causes harm. On the other side, injunctions are issued after an adversarial procedure takes place.[64] There are two significant impacts from this type of restriction that have been considered to be liable to cause a chilling effect that discourages speakers from their communicative act: first, it creates a burden of initiative on the speaker, who has to apply for licenses or permits or file to vacate or modify the terms of an injunction, and, second, it also entails a delay in the communicative act.[65] Moreover, beyond the potential chilling effects that they may cause, prior restraints have also been considered to be problematic because they constitute an 'adjudication in the abstract', meaning that the assessment of publications is done before any actual consequences manifest themselves, which involves some degree of speculation and presumptions.[66]

In some jurisdictions outside Europe, prior restraints are considered so dangerous that they are either banned or nearly impossible to apply. A notable example is Article 13 of the American Convention on Human Rights, which, influenced by U.S. doctrine, prohibits prior censorship,

[48] Mohit Agrawal and R Leela Velusamy, 'R-SALSA: A Spam Filtering Technique for Social Networking Sites', *2016 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)* (IEEE 2016) <http://ieeexplore.ieee.org/document/7509326/> accessed 5 June 2025; G Shanmugasundaram, S Preethi and I Nivedha, 'Investigation on Social Media Spam Detection', *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (2017) <https://ieeexplore.ieee.org/document/8275931> accessed 5 June 2025.

[49] Romero Moreno (n 28).

[50] Mohit Singhal and others, 'SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice', *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)* (IEEE 2023) <https://ieeexplore.ieee.org/document/10190527/> accessed 19 July 2024.

[51] Gorwa, Binns and Katzenbach (n 30).

[52] Google, 'Fighting Child Sexual Abuse Online' <https://protectingchildren.google/> accessed 17 February 2025.

[53] X, 'X's DSA Transparency Report - October 2024' <https://transparency.x.com/dsa-transparency-report.html> accessed 3 March 2025.

[54] TikTok, 'TikTok's DSA Transparency Report July-December 2024' (TikTok 2024) <https://www.tiktok.com/transparency/en/dsa-transparency/>.

[55] Meta, 'Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook' (Meta 2024) <https://transparency.meta.com/sr/dsa-transparency-report-sep2024-facebook>; Meta, 'Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram' (Meta 2024) <https://scontent-arn2-1.xx.fbcdn.net/v/t39.8562-6/466943155_1291701138400105_7867447844898917200_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=l4hNYB1WguUQ7kNvgFY3XUX&_nc_oc=AdhsV902YotvbljHEDBo2zjMeh6cAeVRNOGCQ3INm72Dh9NW7wAM4p1AzbHrLAWjQN8&_nc_zt=14&_nc_ht=scontent-arn2-1.xx&_nc_gid=AguzRsLZO9UkGrkML7wQyxJ&oh=00_AYCDlPUiODY7a3VEg6cgdv1OBpAo3CBSwhGeskeY7pJvpA&oe=67CFB15D>.

[56] Google, 'EU Digital Services Act (EU DSA) Biannual VLOSE/VLOP Transparency Report' (Google 2025) <https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2024-7-1_2024-12-31_en_v1.pdf>.

[57] *Wille v Liechtenstein* App no 28396/95 (ECtHR, 28 October 1999), para 43.

[58] *Altuğ Taner Akçam v Turkey* App no 27520/07 (ECtHR, 25 October 2011), para 81.

[59] *Republic of Poland v European Parliament and Council of the European Union* (n 3), paras 53-55.

[60] *Melike v Turkey* App no 35786/19 (ECtHR, 16 June 2021), para 39.

[61] Michael I Meyerson, 'The Neglected History of the Prior Restraint Doctrine: Rediscovering the Link Between the First Amendment and the Separation of Powers' (2001) 34 Indiana Law Review 295.

[62] Thomas I Emerson, 'The Doctrine of Prior Restraint' (1955) 20 Law and Contemporary Problems 648.

[63] Vincent Blasi, 'Toward a Theory of Prior Restraint: The Central Linkage' (1981) 66 Minn. L. Rev. 11.

[64] ibid.

[65] ibid.

[66] ibid.

allowing only subsequent liability.[67] In Europe, the ECtHR addressed cases related to prior restraints as early as the 1970s[68] but the first time in which it explicitly examined their impact on freedom of expression was in the 1991 case of *Observer and Guardian v. the UK*. Rejecting the position held by international NGO Article 19 advocating for the U.S. approach, the Court held that prior restraints are not expressly prohibited by Article 10 ECHR, but their risks require 'the most careful scrutiny.'[69] This divergence in how the ECHR approaches prior restraints is central to the analysis in this article. Although major online platforms were developed under the U.S. system, which treats prior restraints with strong scepticism, operating in the EU requires compliance with a framework where such measures may be permissible in narrowly defined cases, subject to procedural safeguards. Notably, the ECtHR has explicitly declined to adopt the U.S. approach to prior restraints.

The Court did not strictly define prior restraints in the aforementioned case, but noted that they are implicitly allowed by Article 10's wording on 'conditions,' 'restrictions,' and 'prevention.'[70] Later case law referred to them as 'preventive restrictions', covering measures taken before a final ruling on a case.[71] For the purposes of this article, and considering upload filters' specific features, prior restraints can be understood as preventive interferences on expression, imposed before such expression takes place or before a final ruling on its legality, potentially involving conditions, limitations, or prohibitions on dissemination.[72]

In *Observer and the Guardian v. the UK*, the Court initially emphasized that the careful scrutiny to be applied to prior restraints is particularly relevant to those imposed on press publications, recognizing that the perishable nature of news means delays can diminish its value and public interest.[73] However, ECtHR case law in the digital era has extended this reasoning to 'publications other than periodicals that deal with a topical issue.'[74] Moreover, another key differentiation brought by the ECtHR when reviewing prior restraints in digital contexts is the inclusion of an assessment of whether 'by rendering large quantities of information inaccessible, they substantially restrict the rights of Internet users and have a significant collateral effect.'[75] This stance was adopted by the ECtHR in *Ahmet Yıldırım v. Turkey*, where it found a violation of Article 10 ECHR after authorities imposed a blanket temporary ban on Google Sites due to one allegedly unlawful page. In his concurring opinion, Judge Pinto de Albuquerque outlined eleven criteria, based on Council of Europe documents, for assessing the compatibility of internet blocking laws with the Convention. These include defining who can be targeted by blocking orders, who may issue them, the types (website, IP addresses, protocols, or certain usages like social media) and the scope of orders (national, regional or global), their duration, the legitimate aims under Article 10(2) justifying them, procedural safeguards, assessment of proportionality and necessity, notification of the order and its grounds, and judicial appeal. He also stressed that measures should not affect those that are not *de jure* or *de facto* responsible for the illegal publication and have not endorsed its content.'[76] Formulated in 2012, which means that they may need to be updated to reflect the evolving dynamics of upload filters on social media, which are shaped by both public and private rule-making, with a heavy emphasis on private enforcement, these criteria still serve as a useful point of reference for evaluating current practices.[77]

The principles set out by the previous case law formed the backdrop for the Grand Chamber's 2015 decision in *Delfi AS v. Estonia*, which addressed the liability of an online news portal, Delfi, for failing to promptly remove offensive user comments from its website. This judgment, however, did not take into consideration the criteria set out by judge Pinto de Albuquerque or the possible ways to adapt them to the particularities of this case. The majority of the Court emphasized that large commercial news portals are better equipped 'to prevent or rapidly remove' hate speech than victims and that States may impose liability on such websites if they fail to promptly remove clearly unlawful content, even without prior notice.[78] Since Delfi had a word-based filter, a reporting mechanism, and content rules, the Court found it had not 'wholly neglected' its duties[79] but noted that the mechanism 'failed to filter out odious hate speech and speech inciting violence', leaving it online for six weeks despite the fact that most of the content 'did not include sophisticated metaphors or contain hidden meanings or subtle threats'.[80]

Early commentary on the case highlighted internal tensions within the Court, as seen in its concurring and dissenting opinions.[81] Three concurring judges warned that liability for failing to 'prevent' unlawful comments would require automated or manual pre-monitoring, disproportionately restricting expression.[82] Two dissenting judges argued that a duty to remove offensive content immediately after being posted, without actual knowledge of their existence, incentivized constant monitoring and preemptive moderation, amounting to a blanket prior restraint.[83] They also criticized the majority for failing to assess the

[67] William A Schabas, *The European Convention on Human Rights: A Commentary* (Oxford University Press, Incorporated 2015) <http://ebookcentral.proquest.com/lib/universitetsbiblioteket-ebooks/detail.action?docID=4310766> accessed 11 February 2025.

[68] *Handyside v the United Kingdom* App no 5493/72 (ECtHR, 7 December 1976); *The Sunday Times v the United Kingdom (no 1)* App no 6538/74 (ECtHR 26 April 1979)

[69] *Observer and Guardian v the United Kingdom* (n 2), para 60.

[70] ibid.

[71] ibid, paras 57 and 64.

[72] It should be noted however, that scholars have criticized the broad use of the term 'prior restraint' as it may turn the concept into an unintelligible concept. See for instance Blasi (n 73); John Calvin Jeffries, 'Rethinking Prior Restraint' (1983) 92 The Yale Law Journal 409. The latter notes that, in the U.S. context, 'Some prior restraints involve permit requirements; others do not. Some involve injunctions; others do not. Some cases involving neither permits nor injunctions are treated as prior restraints; others are not. The doctrine purports to deal with matters of form rather than of substance, but there is no unity among the forms of government action condemned as prior restraints.'

[73] *Observer and Guardian v the United Kingdom* (n 2), para 60.

[74] *Ahmet Yildirim v Turkey* App no 3111/10 (ECtHR, 18 December 2012), para 47.

[75] ibid, para 66.

[76] *Ahmet Yildirim v Turkey* App no 3111/10 (ECtHR, 18 December 2012) (Concurring Opinion of Judge Pinto de Albuquerque).

[77] Gonenc Gurkaynak, Ilay Yilmaz and Derya Durlu, 'Exploring New Frontiers in the Interface between Free Speech and Access Bans - The ECHR's Case of Ahmet Yıldırım v. Turkey' (2014) 5 European Journal of Law and Technology <https://ejlt.org/index.php/ejlt/article/view/282> accessed 30 July 2025; Caterina Di Costanzo, 'The Unique Case of Turkey: The European Unions Method in the Quest of Fundamental Rights' (2016) 22 European Public Law 355; Adelina-Maria Tudurachi, 'Internet Access as a Basic Human Right: An Ongoing European Legal Debate?' (2024) 2024 ELTE Law Journal 61.

[78] *Delfi as v Estonia* App no 64569/09 (ECtHR, 16 June 2015), paras 158-159.

[79] ibid, paras 155-156.

[80] ibid, para 156.

[81] Dirk Voorhoof, 'Delfi AS v. Estonia: Grand Chamber Confirms Liability of Online News Portal for Offensive Comments Posted by Its Readers' (*Strasbourg Observers*, 18 June 2015) <https://strasbourgobservers.com/2015/06/18/delfi-as-v-estonia-grand-chamber-confirms-liability-of-online-news-portal-for-offensive-comments-posted-by-its-readers/> accessed 12 February 2025; 'The Delfi AS vs Estonia Judgement Explained' (*Media@LSE*, 16 June 2015) <https://blogs.lse.ac.uk/medialse/2015/06/16/the-delfi-as-vs-estonia-judgement-explained/> accessed 12 February 2025; Neville Cox, 'Delfi v. Estonia: Privacy Protection and Chilling Effect' [2015] Verfassungsblog: On Matters Constitutional <https://intr2dok.vifa-recht.de/receive/mir_mods_00001200> accessed 12 February 2025.

[82] *Delfi as v Estonia* App no 64569/09 (ECtHR, 16 June 2015) (Concurring Opinion of judges Raimondi, Karakas, De Gaetano and Kjølbro).

[83] *Delfi as v Estonia* App no 64569/09 (ECtHR, 16 June 2015) (Dissenting Opinion of judges Sajó and Tsotsoria).

quality of the filtering mechanism, merely assuming it should have been simple but had failed, without deeper analysis on issues such as whether the system used was state-of-the-art, whether platforms should be required to implement the most advanced filtering technologies, and whether liability should still apply if they had done so.[84]

Less than a year later, the Fourth Section of the ECtHR, which included both the dissenting judges and one of the concurring judges from *Delfi AS v. Estonia*, delivered a judgment incorporating some of these concerns by reasoning that assuming that companies should expect that some unfiltered comments would be in breach of law would require 'excessive and impracticable forethought capable of undermining freedom of the right to impart information on the Internet.'[85] While the Court has not addressed these principles in relation to the liability of social media platforms, it has noted that the commercial scale of services may allow for more stringent obligations in the hands of intermediaries.[86]

2.3. The characteristics of upload filters as prior restraints

There are several factors relevant to the classification of upload filters as prior restraints. When looking into this, it is important to note that the ECtHR has noted that regulations on the Internet 'undeniably have to be adjusted according to the technology's specific features in order to secure the protection and promotion of the rights and freedoms concerned.'[87] In other words, while upload filters may share some crucial similarities with offline prior restraints, their classification within that category should also pay regard to the specific features of that technology and the different contexts in which it is applied, such as social media.

First, as upload filters are a type of *ex ante* moderation, the outcome of their screening may result in different types of interference with content. Under the DSA's definition, moderation includes '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.'[88] In that sense, upload filters would fall under the definition given to prior restraints in the previous sub-section, namely preventive interferences on expression before it takes place and involving different types of conditions, limitations, or prohibitions on dissemination. In addition to this, it has been argued that the fact that they constitute a restriction on content before a judicial decision is taken is what turns them into prior restraints.[89]

Second, upload filters are similar to traditional prior restraint mechanisms in the sense that they can include a delay element. They are different from an administrative licensing body but closer to judicial proceedings for injunctions on publications. Instead of applying for permission to publish, users are uploading content without necessarily expecting it to be restricted in any way, but submitted to pre-screening to spot potential violations and to decide on the measure to apply against them. Automated tools are self-executing and work immediately,[90] which means that there would not be any delay in the adjudication process. However, a separate issue arises when they erroneously flag unharmful or legitimate content as harmful or illegal, which must go through an appeals process to 'vacate' the filter's decision.[91] During that procedure, delays may arise and, regardless of their length, issues under the ECtHR's case law would ensue when the decision is to disallow content as the view is that even 'a short period' may deprive a publication 'of all its value and interest'.[92] Similarly, delays may also arise if the automated filter acts as a form of triage for humans making the final decision on the validity of publications before making them available.[93] A result of this is that the public debate misses different voices and perspectives, making it less plural.[94]

Third, upload filters resemble traditional prior restraints by placing the burden of initiative on users who wish to challenge filtering decisions, similar to those seeking to vacate injunctions. Users must activate appeals processes and provide arguments, which may be difficult due to a lack of knowledge about freedom of expression safeguards.[95] Since these processes are part of platforms' private systems, they vary across platforms, leading to uncertainty and requiring specialisation.[96] Unlike *ex post* moderation, where content is initially available and not subject to interferences, in the context of upload filters, unless and until the burden of initiative is taken, the content would have always been subject to an interference (e.g., removal, demonetization, labeling, or demotion). Actions not entailing removal, such as demonetization or demotion, also remain in effect until a successful appeal reverses them, meaning that they also bear a burden of initiative.

Fourth, as De Gregorio has argued, the way in which platforms govern their digital spaces within Europe is of 'autonomous quasi-public functions without the need to rely on the oversight of a public authority' for determining the legality of speech.[97] Under the ECHR framework, the State has a duty not to unduly censor speech alongside a positive obligation to implement protective measures 'in the sphere of relations between individuals'.[98] In that sense, the fact that upload filters operate in the private sphere does not remove protections against them in the ECHR framework in the sense that the State must establish adequate safeguards to protect against undue interference on freedom of expression in the private sphere.

Fifth, and in connection to the last point, upload filters are measures that would require a degree of active monitoring by these private bodies. According to Frosio, in principle, 'to filter unwanted content, all content must be monitored'.[99] Looking at this in more detail, at least at EU level, there can be different levels of monitoring, such as: i) general screening of all the content to find all sorts of illegal material; and ii) general targeted screening, aiming to identify specific types of illegal content

[84] ibid, para 36.

[85] *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v Hungary* App no 22947/13 (ECtHR, 2 February 2016), para 82.

[86] *Pihl v Sweden (déc)* App no 74742/14 (ECtHR, 7 February 2017)

[87] *Editorial Board of Pravoye Delo and Shtekel v Ukraine* App no 33014/05 (ECtHR, 5 May 2011).

[88] DSA, art 3(t).

[89] Nunziato (n 30).

[90] James Grimmelmann, 'Regulation by Software' (2005) 114 Yale Law Journal 1719.

[91] Elkin-Koren (n 30).

[92] *Observer and Guardian v the United Kingdom* (n 2), para 60; *Ahmet Yildirim v Turkey* (n 41), para 47.

[93] Llansó (n 30).

[94] Elkin-Koren (n 30).

[95] Blasi (n 73).

[96] Jennifer M Urban, Brianna L Schofield and Joe Karaganis (eds), 'Notice and Takedown in Everyday Practice'.

[97] Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press 2022) <https://www.cambridge.org/core/product/identifier/9781009071215/type/book> accessed 25 March 2025, p 81.

[98] *Appleby and Others v the United Kingdom* App no 61080/14 (ECtHR, 24 September 2003), para 39; See in the same sense *Fuentes Bobo c Espagne* App no 39293/98 (ECtHR, 29 February 2000), para 38, where the Court stated: 'l'article 10 s'impose non seulement dans les relations entre employeur et employé lorsque celles-ci obéissent au droit public mais peut également s'appliquer lorsque ces relations relèvent du droit privé [...]. En outre, dans certains cas, l'Etat a l'obligation positive de protéger le droit à la liberté d'expression contre des atteintes provenant même de personnes privées [...]. En conséquence, la Cour estime que la mesure litigieuse constituait une ingérence dans l'exercice de son droit à la liberté d'expression protégé par le paragraphe 1 de l'article 10.'

[99] Frosio (n 28).

previously identified by courts, other authorities or by third-parties.[100] Other variations may include limited monitoring, such as random checks, targeted monitoring of specific users or IP addresses suspected of illegal activity, or time-restricted monitoring aimed at detecting specific content.[101] In any case, monitoring involves data collection and surveillance of users, which is seen as an interference with privacy, and may also cause a chilling effect.[102] For that reason, monitoring without adequate limitations and safeguards to its proportionality can turn into a violation to privacy and freedom of expression.[103]

Sixth, although it is possible to argue that moderation by social media platforms stems from the exercise of freedom of expression by these companies, under Article 10 this right entails 'duties and responsibilities'. Although they have not yet been dealt with at the ECtHR, that Court has recognised that social media networks 'necessarily have certain obligations'.[104]

Finally, upload filters have a heightened potential risk of collateral censorship in comparison to traditional non-digital prior restraints. It could be argued that the risks of filters would be lower if companies use technologies with low error rates. However, given the large scale of content published online, even low percentages in error rates can mean millions of publications being affected.[105] For example, Meta's Enforcement Report for October–December 2024 states that the company took action on 66.6 million pieces of content under its policy on adult nudity and sexual activity on Facebook. Of those, 3.1 million were appealed, and 1.39 million were ultimately restored, 92,700 without an appeal and 1.3 million following an appeal.[106] This issue of scale arises in both ex ante and ex post applications of automated tools, but is more acute in the former, as it can suppress millions of publications before they ever see the light of day,[107] with the likelihood of causing a 'significant collateral effect' that would surely be unacceptable under an ECtHR lens.[108] This problem reflects a broader dilemma: algorithmic moderation failures can result in both over-removal of legitimate content and over-allowance of harmful material.[109] On one side, for example, one academic study showed that transgender and Black users are disproportionately affected by removals, which often target content that expresses identity or falls into grey areas.[110] This raises discrimination concerns under Article 14 ECHR and Article 21 of the Charter, and threatens pluralism, which the ECtHR has underpinned as a democratic value linked to the respect for cultural, ethnic, and ideological diversity.[111] An example on the other side relates to reports showing that Facebook's moderation failures contributed to the spread of content linked to the Rohingya genocide in Myanmar.[112] In that sense, the ECtHR has stressed that freedom of expression does not protect calls to violence or hate[113] and that States have a positive obligation to ensure a safe space for all to participate in public debate without fear.[114] An inadequate approach to either side would run counter to the DSA's stated objective of ensuring a 'safe, predictable and trusted online environment', as articulated in Article 1(1) of the Regulation.

3. Upload filters under EU secondary legislation

On 26 March 2019, ahead of the closure of the debate on the largest copyright reform the EU had seen in nearly twenty years, the Vice President of the Commission, Andrus Ansip, took the microphone to state: 'Many speakers spoke about upload filters today. To be clear, upload filters are not mentioned in the text of the directive, but as we all know they are already used by big platforms. Voting down the directive will not take away upload filters. The directive will, on the contrary, give our citizens of the [*sic*] right to ask for reuploads and contest removals of content that they should be able to upload.'[115] This moment was the culmination of a complex process, marked by intense lobbying from big tech, citizen protests, and skepticism from the UN FoE Rapporteur over the prospect of mandating upload filters to prevent copyright infringements.[116]

Despite being a turning point in the legislative debates around upload filters, the Copyright Directive is not the sole piece of legislation concerning this issue. Upload filters have been debated at the EU level for years, with the stance gradually shifting from a relatively skeptical position to a more accepting approach. Within that context, there has been a progressive departure from the principle of 'no monitoring obligation' towards responsibility regimes in which proactive monitoring can become relevant, if not mandatory, to avoid liability of platforms for third party content.[117] Interestingly, during the legislative process of the DSA, the LIBE committee proposed banning hosting providers from using 'ex-ante control measures based on automated tools or upload-filtering,' except for detecting bots or content previously classified as manifestly illegal by a judicial authority or qualified staff. This exception required sufficiently reliable technology that minimizes errors and does not block legal content. The committee also proposed mandatory human review of removal decisions by upload filters, with

[100] Angelopoulos and Senftleben (n 32).
[101] Urban, Schofield and Karaganis (n 106).
[102] 'Algorithms and Human Rights - Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications' (*Council of Europe Publishing*) <https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html> accessed 14 February 2025.
[103] Heldt (n 5).
[104] *Sanchez v France* (n 14), para 185. This point was also raised by the ECtHR in the recent case of Google LLC and others v. Russia. See: *Google LLC and others v Russia* App no 37027/22 (ECtHR, 8 July 2025)
[105] This argument was used recently by Meta to reduce its reliance on automation in the US, although the reference was to automation in general and not necessarily upload filters. See Meta, 'More Speech and Fewer Mistakes | Meta' (7 January 2025) <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/> accessed 18 January 2025.
[106] 'Community Standards Enforcement | Transparency Center' <https://transparency.meta.com/reports/community-standards-enforcement/adult-nudity-and-sexual-activity/> accessed 31 March 2025.
[107] Llansó (n 30).
[108] *Ahmet Yildirim v Turkey* (n 41), para 66.
[109] Vargas Penagos, 'ChatGPT, Can You Solve the Content Moderation Dilemma?' (n 45).
[110] Oliver L Haimson and others, 'Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 1.
[111] *Bączkowski and Others v Poland* App no 1543/06 (ECtHR, 24 September 2007), para 62.
[112] Amnesty International, 'Myanmar: Facebook's Systems Promoted Violence against Rohingya; Meta Owes Reparations – New Report' (*Amnesty International*, 29 September 2022) <https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/> accessed 16 June 2025.
[113] *Dilipak v Turkey* App no 29680/05 (ECtHR, 15 September 2015), para 62.
[114] *Khadija Ismayilova v Azerbaijan* Apps no 65286/13, 57270/14 (ECtHR, 10 January 2019), para 158.
[115] 'Verbatim Report of Proceedings - Copyright in the Digital Single Market (Debate) - Tuesday, 26 March 2019' (n 3).
[116] Emmanuel Vargas Penagos, 'La directiva europea de derechos de autor en el Parlamento Europeo: lobby, protestas y narrativas' (Linterna Verde 2019) <https://www.linternaverde.org/guias/la-directiva-europea-de-derechos-de-autor-en-el-parlamento-europeo> accessed 5 March 2025.
[117] Frosio (n 32).

moderators trained in legal and human rights standards, ensuring fair, transparent, and non-discriminatory moderation.[118] However, the final version of the DSA did not adopt this position and does not make explicit reference to upload filters. Instead, upload filters are allowed or mandated in different layers of the existing legislation.

In this regard, this section will first examine whether upload filters can be mandated under EU law. It will then explore whether platforms are allowed to voluntarily implement upload filters and, if so, what rules govern their application within the EU legal framework.

3.1. The rules for mandatory upload filtering

3.1.1. The general rules on mandatory upload filtering

For over twenty years, the main rules for upload filtering were established in the 2000 E-Commerce Directive, particularly those related to internet intermediaries' liability for third-party content.[119] Following that moment, several pieces of CJEU case law have shaped the interpretation of these rules throughout the years. These are rules that establish a 'safe harbour' for online platforms to avoid liability for the illegality of content uploaded by their users to their services. It allows them to avoid being 'exposed to most forms of liability in the domestic law of the Member States, if they behave in a prescribed way.'[120] To avoid repetition, and especially because the DSA replicates the wording and sense of this part of the E-Commerce Directive, this section does not explain the text of the latter in detail. Instead, it is centred on the DSA, which applies as the general framework since February 2024.[121]

The existing rules include a differentiation between three categories of internet providers: those engaged in 'caching', 'mere conduit' and 'hosting' providers. The two first categories cover services that only transmit or temporarily store content and therefore are subject to minimal and technical rules for liability of third party content.[122] 'Hosting providers', which include social media platforms, app stores and marketplaces, are subject to a stricter regime.[123] They are exempt from liability for third-party content only if they lack actual knowledge of illegal activities or content on their services. However, once they become aware of such content, they must act expeditiously to remove it or disable access to maintain their exemption. In addition to this, the CJEU has established that, in order to retain these protections, platforms must play a neutral role in relation to the illegal content, meaning that 'its conduct is merely technical, automatic and passive, pointing to a lack of knowledge or control of the data which it stores.'[124]This, at the same time, means that platforms would be considered as having actual knowledge if they play an 'active role'.[125] In that sense, the Court has considered that online sales platforms can be considered as playing an active role if they provide 'assistance which entails, in particular, optimising the presentation of the offers for sale in question or promoting those offers.'[126] The CJEU has also specified that implementing 'technological measures aimed at detecting' illegal content does not turn the role of platforms into an active one.[127] Furthermore, the DSA introduced a provision clarifying that 'actual knowledge' can be obtained through notices from users.[128]

These rules for immunity for third-party content are coupled with a prohibition on Member States imposing general monitoring obligations,[129] which could be seen as a safeguard against mandatory filtering. Such a mandate, initially provided for in Article 15 of the E-Commerce Directive, and now incorporated in Article 8 of the DSA, has been one of the most central points of academic debate and litigation at the CJEU.

In that vein, when interpreting this prohibition, the CJEU has considered that EU law cannot require platforms to implement filtering mechanisms that actively, indiscriminately and preventively monitor all user-stored content, solely at the platform's expense, for unlimited periods, to detect musical, cinematographic, or audiovisual works.[130] According to the CJEU, Member States cannot impose monitoring obligations requiring permanent and costly automated systems that could interfere with the platform's freedom to conduct business protected under article 16 of the EU Charter.[131] In that same line, the Court has also noted that authorities must put into balance the rights of the party affected by the content (e.g. intellectual property), the rights of the publisher (e.g. freedom of expression), the right of the platform to conduct its business, and the rights of the platform's users, such as privacy and freedom to receive or impart information.[132]

Additionally, the CJEU has established that any filtering mechanisms must be capable of distinguishing between lawful and unlawful content because the opposite would turn into an unacceptable interference to freedom of expression.[133] In this context, measures directed at specific users or sellers engaging in repeated infringements are permitted, as long as they balance effectiveness and deterrence without creating barriers to legitimate trade and do not require active monitoring of all platform activity.[134]

Furthermore, the CJEU has decided that national courts are allowed

[118] Patrick Breyer, 'OPINION on the Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and Amending Directive 2000/31/EC | LIBE_AD(2021)692898 | European Parliament' <https://www.europarl.europa.eu/doceo/document/LIBE-AD-692898_EN.html> accessed 22 March 2025.

[119] 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 [2000] OJ L 178, 17.7.2000, pp. 1–16 (E-Commerce Directive), former arts 14 and 15.

[120] Martin Husovec, 'Holey Cap! CJEU Drills (yet) Another Hole in the e-Commerce Directive's Safe Harbours' (2017) 12 Journal of Intellectual Property Law & Practice 115.

[121] In that sense, Recital 16 of the DSA notes that the regime provided by the E-Commerce directive 'has allowed many novel services to emerge and scale up across the internal market. That framework should therefore be preserved.'

[122] Recital 29 of the DSA provides examples of those services, such as wireless access points, virtual private networks, or voice over IP calls for 'mere conduit' and content delivery networks, reverse proxies or content adaptation proxies for 'caching'. For a more detailed analysis of the regime of these services, see: Sebastian Felix Schwemer, Tobias Mahler and Håkon Styri, 'Liability Exemptions of Non-Hosting Intermediaries: Sideshow in the Digital Services Act?' (2021) 8 Oslo Law Review 4.

[123] E-Commerce Directive, former Art 14; DSA, Art 6.

[124] Joined Cases C-236/08 to C-238/08 *Google France SARL, Google Inc v Louis Vuitton Malletier SA; Viaticum SA,Luteciel SARL, and Google France SARL v Centre national de recherche en relations humaines (CNRRH) SARL, Pierre-Alexis Thonet, Bruno Raboin, Tiger SARL* [2010] EU:C:2010:159 (Google France), para 114 See also Case C-324/09 *L'Oréal SA, Lancôme parfums et beauté & Cie SNC, Laboratoire Garnier & Cie, L'Oréal (UK) Ltd v eBay International AG, eBay Europe SARL, eBay (UK) Ltd, Stephen Potts, Tracy Ratchford, Marie Ormsby, James Clarke, Joanna Clarke, Glen Fox, Rukhsana Bi* [2011] EU:C:2011:474 (*L'oreal*), para 113.

[125] *Google France* (n 134), para 120; *L'oreal* (n 134), para 113.

[126] *L'Oréal* (n 134), para 116.

[127] Joined Cases C-682/18 and C-683/18 *Frank Peterson v Google LLC, YouTube LLC, YouTube Inc, Google Germany GmbH (C-682/18) and Elsevier Inc v Cyando AG (C-683/18)* [2021] EU:C:2021:503 (*Cyando*), para 109.

[128] DSA, art 16(3).

[129] E-Commerce Directive, former art 15; DSA,art 8.

[130] Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2011] EU:C:2011:771 (Scarlet Extended), paras 29, 36 and 54; *Case C-360/10 Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV,* [2012] EU:C:2012:85 (*Netlog*), paras 34-38; *Cyando* (n 137), para 135.

[131] *Scarlet Extended* (n 140), paras 29 and 54; *Netlog* (n 140), paras 26 and 52.

[132] *Scarlet Extended* (n 140), para 53; *Netlog* (n 140), para 51.

[133] *Scarlet Extended* (n 140), para 52; *Netlog* (n 140), para 50.

[134] *L'Oréal* (n 134), paras 139-140.

to, once an infringement has been determined, issue court injunctions to prevent future infringements as long as they are targeted, 'fair and proportionate', and not 'excessively costly'.[135]

In addition to this, relying on Recital 47 of the E-Commerce Directive, which states that the prohibition on general monitoring does not apply to monitoring obligations 'in a specific case,' the CJEU has established that platforms can, in some cases, be required by court injunctions to prevent the dissemination of content that is not only 'identical' to material previously declared unlawful but also has an 'equivalent meaning,' 'irrespective of who requested the storage of that information.'[136]

These developments from the CJEU have expanded the interpretation of 'no general monitoring' obligation into one in which 'the generality or specificity of the monitoring is not determined by what is being monitored, but by the objective of the monitoring'.[137] This is problematic because it implies that, even if it is just to find specific infringing content, all content must pass through the filtering mechanism.[138]

To ensure balance, the CJEU also established that 'equivalent information' must be properly identified in the injunction ordering the platform to prevent its dissemination. This includes details such as 'the name of the person concerned by the infringement determined previously, the circumstances in which that infringement was determined, and equivalent content to that which was declared to be illegal.'[139] Furthermore, the CJEU has stated that platforms should not be required to 'carry out an independent assessment' to determine the equivalence of content.[140] In this context, 'independent assessment' refers to platforms being tasked to make their own legal judgments on the content, something that the CJEU considers as avoided when platforms have 'recourse to automated search tools and technologies.'[141] In that sense, the Court has given preference to the use of automated filters in these circumstances without giving details of what characteristics would make the tool adequate, something that can be problematic because filters may present difficulties determining lawful uses of those potential re-infringements, for example journalistic or academic commentary.[142] Additionally, by saying that the use of automated tools is enough to consider that the platform is not conducting an 'independent assessment' seems to ignore that platforms are the ones who must introduce the criteria used by the filters to identify content. This is a process involving several legal and technical interpretations and decisions by the platform in the design and programming of the filter. Despite not making a case-by-case direct analysis, the way in which platforms exercise control over the filter makes them in charge of carrying out their own assessment on how to interpret and operationalize the injunction.

While the CJEU's interpretation of the prohibition on general monitoring obligations remains unchanged, it remains to be seen whether Recital 30 of the DSA may influence a change of perspective. Recital 30 points to the allowance of monitoring obligations monitoring in specific cases and compliance with national orders under EU law, but introduces nuances that might influence future interpretation by the CJEU. It states that platforms should not be subject to general monitoring obligations, either *de jure* or *de facto*, nor to general proactive duties to address illegal content. While its wording does not depart from earlier complications introduced by the CJEU, it may support arguments for narrowing monitoring duties in the future.

These possibilities for imposing orders for specific monitoring are partly reflected in Article 9 DSA, which establishes the possibility for administrative or judicial authorities to enact orders for platforms 'to act against one or more specific items of illegal content'. These orders must establish their legal basis, an explanation of why the content is deemed to be illegal, the issuing authority, details of where to locate the content (such as URLs), information on redress mechanisms available, their territorial scope, and, where applicable, details on the authority monitoring compliance. Additional standards for orders for the removal of certain content constituting violence against women were recently introduced by the VAW Directive.[143]

3.1.2. Article 17 of the copyright directive as a lex specialis provision on mandatory upload filtering

The Copyright Directive establishes a liability regime under which platforms qualifying as 'online content-sharing service providers' (OCSSPs) can be held responsible for copyright infringement committed by their users unless they implement a series of 'best efforts' against such infringement. OCSSPs are platforms operating with the main purpose of storing and providing public access to large amounts of copyrighted content that's uploaded by their users, which they organize and promote for profit.[144] Senftleben et al. note that, although relatively unclear in wording, 'it is safe to assume that certain large-scale platforms, especially platforms with video-sharing features (e.g., YouTube, Facebook, Instagram)' fall under this definition.[145] The regime established under Article 17 of this Directive mandates that platforms obtain authorization, such as licensing agreements, from copyright holders to avoid liability for providing access to content uploaded by their users. For instance, if a user uploads a video featuring a song by Shakira without being Shakira or someone authorized by her, the platform may be held liable for an unauthorized communication to the public of that song, unless the platform itself has secured a license for the content in question. If the platform lacks a license, it can avoid liability by making 'best efforts' and following 'high industry standards of professional diligence' to 'ensure the unavailability' of works for which rightholders have provided 'relevant and necessary information,'[146] as well as to prevent their future uploads.[147] For example, if Shakira or her representatives have previously identified certain songs as material that should be made unavailable due to a lack of authorization, and a user later attempts to upload one of those songs without permission, the platform must apply 'best efforts' to ensure that the content is not made available. Commentary on this part of the Directive has noted that, although it does not explicitly include the Commission's original proposal to apply 'content recognition technologies,' its wording can only be interpreted as requiring upload filtering mechanisms.[148] Regardless of the technology's good intention to protect copyright, it can be subject to abuse. For instance, in 2021, media reports described how a police officer began

[135] Ibid, paras 139-140.
[136] Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] EU:C:2019:821 (*Eva Glawischnig-Piesczek)*, paras 34 and 53.
[137] Angelopoulos and Senftleben (n 32).
[138] Frosio (n 28).
[139] *Eva Glawischnig-Piesczek* (n 146), para 46.
[140] ibid, para 45.
[141] ibid, para 45.
[142] Daphne Keller, 'Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling' (2020) 69 GRUR International 616.
[143] VAW Directive, art 23.
[144] Copyright Directive, art 2(6).
[145] Senftleben, Martin; Quintais, João Pedro; Meiring, Arlette; (n 33).
[146] Copyright Directive, art 17(4)(b).
[147] ibid, art 17(4)(c).
[148] Joao Pedro Quintais and others, 'Safeguarding User Freedoms in Implementing Article 17 of the Copyright in the Digital Single Market Directive: Recommendations from European Academics' (2020) 10 Journal of intellectual property, information technology and electronic commerce law 277; Romero Moreno (n 28); Senftleben, Martin; Quintais, João Pedro; Meiring, Arlette; (n 33).

playing 'Blank Space' by Taylor Swift aloud while he was, at the same time, conducting a proceeding to allegedly limit a demonstration for police reform in Oakland. When questioned by activists, the officer responded, 'You can record all you want. I just know it can't be posted to YouTube.'[149] Beyond the cleverness of the police officer's strategy, his actions can be seen as limiting the capacity of the demonstrators to post information and opinions about his conduct as an official, an issue that has been considered a matter of public interest.[150]

The CJEU examined the validity of Article 17 of the Copyright Directive in an action for annulment filed by the Polish government. It concluded that, indeed, the wording of that provision obliges platforms to conduct prior reviews of content that 'depending on the number of files uploaded and the type of protected subject matter in question' may require the use of upload filters.[151] The CJEU saw this requirement as a limitation of rights under Article 11 of the Charter,[152] but considered that it was not disproportionate. It noted that the obligation for preventing availability of unauthorised content is merely one of 'best efforts' while the obligation of protecting lawful content 'prescribes a specific result to be achieved.'[153] This is complemented by non-binding guidance from the Commission, mandated by Article 17(10), which states that platforms only have to screen for 'manifestly infringing' content.[154] Furthermore, the CJEU also underpinned that Article 17 included safeguards allowing that lawful content remains available, including mandatory exceptions for quotation, criticism, and parody,[155] the absence of a general monitoring obligation,[156] and remedial safeguards explained in Section IV. Against this background, and based on the AG opinion in this case, there is commentary proposing that content not 'manifestly infringing' would be subject to a 'presumption of lawfulness' and that platforms must not limit themselves to afford *ex post* safeguards (i.e. complaints, redress, review) but also *ex ante* measures, for example limiting the application of unreliable filtering technologies that fail in conducting contextual assessments.[157]

Article 17 has a less stricter regime, which would not demand upload filtering, for relatively young and small platforms with less than three years of existence, with annual turnover below EUR 10 million and average monthly unique visitors below five million. However, since these companies will eventually be required to fully comply with Article 17, they will be at a competitive disadvantage compared to big tech companies with greater financial resources, something that may, at the same time, push them to implement upload filters from an early point in their existence.[158]

Academic commentary has emphasized that this article complements the DSA's obligations.[159] Quintais and Schwemer note that Article 17(3) of the Copyright Directive creates a separate liability regime for copyright protection, exempting it from the DSA's liability regime. Furthermore, Article 8 DSA's preclusion on general monitoring remains intact. Beyond that, as those authors note, both regimes are complementary in the sense that the DSA offers horizontal rules that address areas not covered by Article 17 and guide Member States where Article 17 allows discretion, mainly through procedural safeguards and foundational principles for intermediary regulation.[160] Senftleben notes that the DSA's adoption was an opportunity to refine the EU approach to protecting human rights affected by algorithmic filtering. However, he considers that there is a 'surprising' reliance on user complaints to activate those safeguards.[161] However, as argued in subSection 4.2, the integration of procedural human rights standards shapes how upload filters are internally implemented within the moderation cycle.

3.2. The rules for voluntary upload filtering

3.2.1. The general rules on 'voluntary' upload filtering

3.2.1.1. Freedom to filter. In Article 7, the DSA provides that platforms are allowed to implement 'voluntary own-initiative' measures to detect, identify and remove illegal content as long as they are done 'in good faith and in a diligent manner'. Furthermore, Quintais and Schwemer argue that Article 7 DSA may complement Article 17 of the Copyright Directive in the sense that voluntary measures may go beyond the 'best efforts' required by the latter, as long as they comply with the safeguards set in Article 17(7)-(9).[162]

Although the purpose of this article is not to provide a comprehensive comparative analysis between EU and US legislation on the subject matter, it should be noted that Article 7 DSA has been seen as the embodiment of the Commission's intention to make an EU transplant of the 'good samaritan' protection given in Section 230(1)(c) of the US Communications Decency Act.[163] However, there is academic commentary at EU level noting the differences between both provisions and signalling possible risks for overcompliance that it may bring.

The US provision allows platforms to take voluntary actions in good faith 'to restrict access to or availability of material' considered 'obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable' by the platform or an user, 'whether or not such material is constitutionally protected'.[164] On the other hand, Husovec notes that Article 7 is just a codification of CJEU's case law that

[149] Zoë Schiffer and Adi Robertson, 'Watch a Police Officer Admit to Playing Taylor Swift to Keep a Video off YouTube | The Verge' (*The Verge*) <https://www.theverge.com/2021/7/1/22558292/police-officer-video-taylor-swift-youtube-copyright> accessed 23 March 2025.

[150] *Toth and Crișan v Romania* App no 45430/19 (ECtHR, 25 February 2025), paras 60-63.

[151] *Republic of Poland v European Parliament and Council of the European Union* (n 12), paras 53-54.

[152] ibid, para 55.

[153] *Republic of Poland v European Parliament and Council of the European Union* (n 12), para 78.

[154] Commission, 'Communication from the Commission to the European Parliament and the Council, Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market' COM/2021/288 final.

[155] *Republic of Poland v European Parliament and Council of the European Union* (n 12), para 87; Copyright Directive, art 17(7).

[156] *Republic of Poland v European Parliament and Council of the European Union* (n 12), para 90; Copyright Directive, art 17(8).

[157] João Pedro Quintais and Sebastian Felix Schwemer, 'The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?' (2022) 13 European Journal of Risk Regulation 191.

[158] Thomas Spoerri, 'On Upload-Filters and Other Competitive Advantages for Big Tech Companies under Article 17 of the Directive on Copyright in the Digital Single Market' (2019) 10 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 173.

[159] Quintais and others (n 158); Quintais and Schwemer (n 167); Alexander Peukert and others, 'European Copyright Society – Comment on Copyright and the Digital Services Act Proposal' (2022) 53 IIC - International Review of Intellectual Property and Competition Law 358; Martin Senftleben, 'Guardians of the UGC Galaxy – Human Rights Obligations of Online Platforms, Copyright Holders, Member States and the European Commission Under the CDSM Directive and the Digital Services Act' (2023) 14 JIPITEC <http://www.jipitec.eu/issues/jipitec-14-3-2023/5847>.

[160] Quintais and Schwemer (n 167).

[161] Senftleben (n 169).

[162] Quintais and Schwemer (n 167).

[163] Aleksandra Kuczerawy, 'The Good Samaritan That Wasn't: Voluntary Monitoring under the (Draft) Digital Services Act' [2021] Verfassungsblog: On Matters Constitutional <https://intr2dok.vifa-recht.de/receive/mir_mods_00009842> accessed 7 March 2025; Aina Turillazzi and others, 'The Digital Services Act: An Analysis of Its Ethical, Legal, and Social Implications' (2023) 15 Law, Innovation and Technology 83.

[164] Protection for private blocking and screening of offensive material 1996 (47 USC 230).

had already noted that companies do not lose neutrality if they take technological measures to detect illegal content.[165]

The 'good samaritan' protection in the US has been considered relevant for platforms as it allows them to 'experiment with proactive measures' without ceasing to be considered 'neutral' when moderating. As such, this rule is seen as an incentive for moderation because 'companies would, on the whole, try harder to weed out bad content if those efforts didn't expose them to legal risk.'[166] However, this provision has also faced criticism because it gives 'nearly unlimited discretion to remove with impunity' any 'objectionable' content as long as it is done in 'good faith', a concept that, at the same time, has been interpreted broadly by US Courts.[167] In the context of Article 7 DSA, Frosio & Geiger have considered that, while this provision clarifies that voluntary measures for content moderation do not affect platforms' liability exemptions, it may also lead to overenforcement and negative effects on freedom of expression.[168] As these authors note, by ensuring legal certainty for platforms, the provision might incentivize excessive content removal, as platforms can avoid liability by taking down content rather than risking legal disputes.

In addition to the commentary by the mentioned authors, it is worthy to note that Article 7 of the DSA differs from Section 230 in several key aspects. First, while Section 230 refers only to restricting access to content, Article 7 encompasses broader actions, including detecting, identifying, and removing content, as well as ensuring compliance with EU and national laws. This broader reach could grant platforms greater discretion in how they identify and manage content. Second, whereas Section 230 applies to 'objectionable content,' Article 7 is limited to 'illegal content,' which requires legal interpretation and could pressure platforms to over-remove content to avoid liability. Additionally, Article 7 does not explicitly protect good faith content removal based on terms of service; this aspect is only suggested in Recital 26.[169] Finally, Article 7 imposes a stricter standard by requiring platforms to act not only in 'good faith' but also 'in a diligent manner.' Recital 26 hints that diligence entails using sufficiently reliable automated tools to minimize errors, which may imply an expectation of state-of-the-art technology and allowing discretion on whether these tools are employed proactively or reactively. Furthermore, as Husovec notes, the DSA does not create an immunity regime but instead ensures that existing liability exemptions remain.[170] He does note, however, that the strong risk-mitigation requirements in hand of VLOPs makes it difficult to consider that measures can actually be considered 'voluntary'.[171] Indeed, Article 35(1)(c) establishes that one of the possible risk-mitigation strategies that VLOPs can deploy includes adapting the 'speed and quality of processing notices related to specific types of illegal content and, where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence.' In that sense, the attempted transplant made by the EU legislature has different implications from those of its source of inspiration, triggering certain risks of overcompliance.

This should, however, be read in line with Article 14(4), which mandates that all restrictions should be applied in a 'diligent, objective and proportionate manner' with due regard to the fundamental rights at stake in each case. The text of this provision raises several debates in legal scholarship, with an ongoing debate on whether it creates a horizontal application of freedom of expression between platforms and their users.[172] However, as Wendel explains, even if this provision is not a legal basis for such horizontal effect of the rights in question, this does not imply that those rights do not have such effect by themselves in this context. Wendel argues that freedom of expression is among those rights that fulfil the criteria of CJEU's case law on the horizontal application of rights as it is sufficient in itself and confers an individual right that can, by its very own nature, imply a corresponding obligation in the hands of another private party. This is reinforced by the fact that, as Wendel also notes, private platforms 'enjoy structural superiority' over their users in the exercise of freedom of expression online.[173] Furthermore, it should be noted that the DSA links the protection of rights with the principle of consumer protection,[174] which under CJEU's case law affords critical importance to bringing balance in a relationship in which 'the consumer is in a weak position vis-à-vis the seller or supplier, as regards both his bargaining power and his level of knowledge'.[175] Having said that, it is also important to note that Article 54 of the DSA establishes that platforms are liable for any damages suffered by users due to infringements of obligations established in this Regulation.

3.2.1.2. Commitments to filter?. The DSA establishes a co-regulatory framework in which the European Commission and the European Board for Digital Services (the Board) are tasked with encouraging platforms to adopt voluntary codes of conduct to support the implementation of that Regulation.[176] Most of the language on the enforcement of these codes of conduct is relatively soft, as the Commission and the Board are only tasked with 'facilitating' their drafting,[177] 'inviting' VLOPs to participate in the process if there are significant risks emerging from their mandatory risk assessments,[178] 'assessing' whether the codes meet the purposes of the DSA,[179] 'monitoring and evaluating' compliance,[180] and 'inviting' platforms to take 'necessary action' in cases of 'systematic failure' to comply.[181] The DSA introduces additional measures to ensure that VLOPs comply with the codes of conduct, including, among others, annual independent audits that evaluate their commitments under these instruments[182] and the establishment of internal compliance offices tasked, among other duties, with ensuring adherence to the codes.[183] Despite this amicable language, as Husovec notes, 'the pull to join is powerful' and the wording seems to indicate the possibility

[165] Martin Husovec, *Principles of the Digital Services Act* (1st edn, Oxford University Press 2024) <https://academic.oup.com/book/58088> accessed 23 March 2025, p 112.

[166] Daphne Keller, 'Internet Platforms: Observations on Speech, Danger, and Money' (Hoover Institution's Aegis Paper Series 2018) 1807 <https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf>.

[167] Nicholas Bradley, 'Something for Nothing: Untangling a Knot of Section 230 Solutions Student Note' (2022) 2022 Cardozo Law Review De-Novo 58.

[168] Frosio and Geiger (n 31).

[169] '[...]it is appropriate to clarify that the mere fact that the providers take measures, in good faith, to comply with the requirements of Union law, including those set out in this Regulation as regards the implementation of their terms and conditions, should not render unavailable the exemptions from liability set out in this Regulation.', DSA, Recital 26.

[170] Husovec (n 175), p 112.

[171] ibid.

[172] Andrea Palumbo, 'A Medley of Public and Private Power in DSA Content Moderation for Harmful but Legal Content: An Account of Transparency, Accountability and Redress Challenges' (2024) 15 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law <https://www.jipitec.eu/jipitec/article/view/412> accessed 26 March 2025.

[173] Mattias Wendel, 'Taking or Escaping Legislative Responsibility? EU Fundamental Rights and Content Regulation under the DSA' in Antje von Ungern-Sternberg (ed), *Content Regulation in the European Union: The Digital Services Act*, vol 1 (Verein für Recht und Digitalisierung eVInstitute forDigital LawTrier(IRDT) 2023) <https://irdt-schriften.uni-trier.de/index.php/irdt/catalog/book/3> accessed 26 March 2025.

[174] DSA, art 1(1).

[175] Case C-34/13 *Monika Kušionová v SMART Capital, as* [2014] EU:C:2014:2189.

[176] DSA, art 45(1).

[177] ibid.

[178] ibid, art 45(2).

[179] ibid, art 45(4).

[180] ibid.

[181] ibid.

[182] ibid, art 37(1)(b).

[183] ibid, art 41(3)(c).

of impacting VLOPs risk-assessment and mitigation obligations.[184]

There are two codes of conduct that have been adopted following the DSA's adoption that are relevant for the scope of this article. One is the 2022 Code of Practice on Disinformation, which does not include direct reference to upload filtering, but some commitments may be vague enough to allow filtering, such as 'avoiding the publishing and carriage of harmful disinformation to protect the integrity of advertising supported businesses'[185] or implementing 'proportionate policies to limit the spread of harmful false or misleading information'.[186] Such vagueness is not complemented with rules limiting the application of filtering measures.[187] The second one is the 2025 Code of conduct on countering illegal hate speech online +.[188] This instrument seems to have a more limited approach, only focusing on the expeditiousness of platforms' processes to respond to notices by users or trusted flaggers.

3.2.2. *Lex specialis provisions on voluntary upload filtering*

Article 7 DSA strongly incentivizes platforms to adopt upload filters, which, along with *lex specialis* provisions, may make them a common compliance practice. This poses challenges to freedom of expression, as users' content is constantly screened for potential limitations. Additionally, common practices can later become regulatory requirements, as seen with Article 17 of the Copyright Directive, which, as Dusollier notes, turned YouTube's voluntary use of ContentID into a legal obligation.[189]

The following paragraphs refer to legal provisions within the EU legal framework that, despite not making the use of upload filters mandatory, are worded in a way that could incentivize their voluntary use.

3.2.2.1. The audiovisual media services directive. The most recent version of the Audiovisual Media Services Directive introduced new obligations for 'video-sharing platform services.' These are platforms which primarily, or through a dedicated service, provide programs or user-generated videos to the general public without editorial control over the content, organizing them through means that may include algorithms or automation.[190] Article 28b of this Directive provides that platforms must 'take appropriate measures' to protect minors from content that may harm their development, and the general public from content inciting violence or hatred, as well as from content related to certain criminal offenses, including terrorism, child sexual abuse material, and racist or xenophobic material.[191] Article 28b(3) explains that 'appropriate' is something to be determined 'in light of the nature of the content in question, the harm it may cause, the characteristics of the category of persons to be protected as well as the rights and legitimate interests at stake' and that, for the protection of minors, 'the most harmful content shall be subject to the strictest access control measures.' It further provides a non-exhaustive list of measures that can be considered 'appropriate', such as flagging mechanisms for users, age verification mechanisms and parental control mechanisms, among others. While this list of measures does not include upload filters, Article 28b(6) allows Member States to impose 'stricter measures' than those enlisted. As Oruç notes, despite the fact that Article 28b says that its application should not disregard the prohibition on general monitoring, it is difficult to establish which measures can be 'stricter' without constituting ex ante moderation.[192] Furthermore, a 2021 report conducted by Deloitte for the Commission's Directorate-General for Communications Networks, Content and Technology on the implementation of the AVMSD noted that platforms are increasingly relying on *ex ante* 'AI-driven filters to detect and filter the most problematic content' despite them not being mandated by this Directive.[193] Moreover, a review conducted by Cole and Etteldorf points out that Member States who introduced 'stricter' measures (Finland, Germany and Sweden) did not focus on the type of measures but instead on expanding the duties of platforms for them to act against content criminally sanctioned under national law.[194]

3.2.2.2. TERREG. Article 3 of TERREG provides that national authorities can mandate platforms to remove 'terrorist content' within one hour of receiving the order to do so. Furthermore, Article 5(4) establishes that national authorities can designate platforms as 'exposed to terrorist content' when they had received at least two of those orders in the previous 12 months. Such designation activates an obligation by platforms to take 'specific measures'. Although the possible measures enlisted in Article 5(2) are focused on *ex-post* review and flagging mechanisms, and Article 5(8) establishes that there's no obligation to use automated tools or to conduct general monitoring, Articles 5(2)(c) and 5(2)(c) allow any other 'mechanisms to increase awareness' and 'any other measure' considered 'appropriate' by the platform 'to address the availability of terrorist content'. Furthermore, Recital 25 of TERREG hints that platforms can deploy automated tools 'if they consider this to be appropriate and necessary to effectively address the misuse of their services for the dissemination of terrorist content.' The final version of TERREG does not include a provision originally proposed by the Commission, which aimed to require platforms to take 'specific proactive measures' for 'preventing the re-upload of content which has previously been removed or to which access has been disabled because it is considered to be terrorist content.'[195]

However, in practice, platforms apply upload filters for compliance with this legislation. This can be seen in transparency reports mandated by Article 7 TERREG, the most recent of them published in February 2025. For instance, Google's report claims they use 'machine learning technologies' to 'detect re-uploads of violative content, in many cases before it is widely viewed by users' ;[196] Meta's reports for Instagram and Facebook say they use 'media matching' to prevent re-uploads that they have previously detected and hashes from GIFCT and note that both

[184] Husovec (n 175).
[185] Code of Conduct on Disinformation 2022, Measure 1.1.
[186] ibid, Measure 18.2.
[187] 'How Will the EU Digital Services Act Affect the Regulation of Disinformation?' (2023) 20 SCRIPTed <https://script-ed.org/?p=4119>.
[188] Code of Conduct on Countering Illegal Hate Speech Online 2025.
[189] Séverine Dusollier, 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition' (2020) 57 Common Market Law Review 979.
[190] Audiovisual Media Services Directive, art 1(1)(b)(aa).
[191] Ibid, art 28b.
[192] Toygar Hasan Oruç, 'The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in Light of Recent Developments: Is It Still Necessary to Maintain It?' (2022) 13 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 176.
[193] European Commission. Directorate General for Communications Networks, Content and Technology., Deloitte., and SMIT., *Study on the Implementation of the New Provisions in the Revised Audiovisual Media Services Directive (AVMSD): Final Report, Part D.* (Publications Office 2021) <https://data.europa.eu/doi/10.2759/135983> accessed 6 March 2025.
[194] Mark D Cole and Christina Etteldorf, *Future Regulation of Cross-Border Audiovisual Content Dissemination: A Critical Analysis of the Current Regulatory Framework for Law Enforcement under the EU Audiovisual Media Services Directive and the Proposal for a European Media Freedom Act* (Nomos Verlagsgesellschaft mbH & Co KG 2023) <https://www.nomos-elibrary.de/index.php?doi=10.5771/9783748939856> accessed 6 March 2025.
[195] Commission, 'Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on preventing the dissemination of terrorist content online' 2018/0331 (COD).
[196] Google, 'Regulation (EU) 2021/784 on Addressing the Dissemination of Terrorist Content Online Transparency Report' (Google 2025) <https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-26_2024-1-1_2024-12-31_en_v1.pdf>.

technologies allow that terrorist 'content intended for upload' 'never reaches' their platforms ;[197] TikTok's report says they use computer vision models, text-based technologies, such as blacklists or NLP tools, and hashing technologies.[198] Although this report does not signal if TikTok uses these technologies *ex ante*, their DSA report mentioned above serves as clarification that they do.[199]

Furthermore, as Coche[200] notes, there is a risk of over censorship in compliance with TERREG because of its broad definition of Terrorist content, as well as the possibility brought by Article 4(7) for platforms to remove content based on their terms of service, even if that content was previously misclassified as illegal. Such discretion in the hands of platforms may allow them to apply filtering for issues beyond what is required by law.

3.3. Filtering of lawful content

The legal framework at EU level is not limited to establishing rules on upload filtering against illegal content. Instead, it allows platforms to establish terms and conditions as the basis for restrictions to user generated content in their services. This is evident in the aforementioned article 4(7) of TERREG, but also in the definition of 'content moderation' given by the DSA, which refers to it as a process aimed at addressing both illegal content and content 'incompatible' with terms of service.

This possibility to restrict content that is legal, but against terms and conditions, has been labelled the possibility of limiting content deemed to be 'lawful but awful'. In other words, content 'offensive or morally repugnant to many people but protected' by speech.[201] This category can also include spam, bots, scams and other types of content that is not necessarily illegal, but platforms need to filter as a way to ensure a better user experience within their services.[202]

Notably, a study by Kaushal et al. analyzing the statements of reasons submitted by VLOPs to the Transparency Database mandated by Article 24(5) and maintained by the European Commission found that 99.8 % of 131 million content restrictions were based on violations of terms and conditions, while only 0.2 % cited illegality as the reason.[203] This matches conclusions of previous scholarship noting that the regulation of online discourse is increasingly centred in private unilateral rulemaking.[204]

In principle, filtering of terms and conditions is meant to be voluntary. However, the due diligence obligations of VLOPs may indirectly turn it into an obligation. In that regard, Article 34(2)(c) of the DSA establishes that the risk assessment to be conducted by these platforms must take into account, among other factors, whether their terms and conditions and their enforcement are contributing to systemic risks such as the dissemination of disinformation, content harmful to rights or public interests, including civic discourse and elections, among others. When identified, Article 35(1)(b) of the DSA mandates the introduction of risk-mitigation strategies, which can include adapting terms and conditions and their enforcement. Going further, these measures can even be considered to be mandatory during a 'crisis' declared by the Commission through a Recommendation of the European Board for Digital Services, based on the existence of 'extraordinary circumstances' that pose a 'serious threat to public security or public health' in accordance with Article 36(1)(b) DSA. Furthermore, as can be seen in subsection 2.1.2, Codes of Conduct may introduce commitments as regards limitations on content that is not illegal, such as disinformation. In that sense, as Palumbo points out, there is a risk that these rules facilitate pressure by public bodies—particularly the Commission—to mandate restrictions on content that is not illegal.[205]

4. Multilayered remedial safeguards against undue upload filtering in EU law

As explained above, the CJEU has embraced the ECtHR's case law in relation to prior restraints for saying that systems in which upload filters are applied require 'a particularly tight legal framework' to avoid encroachments on freedom of expression.[206] Furthermore, the emphasis given by the ECtHR is that any such framework must ensure 'tight control over the scope of bans and effective judicial review to prevent any abuse of power',[207] as well as procedural safeguards against arbitrary restrictions to that right.[208] In addition to this, when analysing the scope of Article 47 of the Charter, the CJEU has considered that, under the principle of effectiveness, 'procedural rules governing actions for safeguarding an individual's rights' must not 'render practically impossible or excessively difficult the exercise of rights conferred by EU law'.[209] The latter has been relevant at EU level even from before the Charter[210] and is enshrined at Article 19(1) TEU, establishing an obligation for Member States to 'provide remedies sufficient to ensure effective legal protection in the fields covered by Union law'.

Moreover, Article 6 ECHR provides that everyone enjoys the right to an impartial and independent 'tribunal' in the determination of their civil rights and obligations, which includes disputes related to contesting interferences against freedom of expression.[211] When referring to the concept of a 'tribunal', the ECtHR has referred to institutions exercising a 'judicial function', namely 'determining matters within its competence on the basis of rules of law and after proceedings conducted

[197] Meta, 'European Union Terrorist Content Online Transparency Report' (Meta 2025) <https://scontent-arn2-1.xx.fbcdn.net/v/t39.8562-6/480671819_1003419394326168 7_547247169033819793_n.pdf?_nc_cat=105&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=Gbo5XUpto3IQ7kNvgGOVHHP&_nc_oc=AdhcvcDm2wC-EsvjiZoZYBi_9Gel5apY0Mk36JnKHBXj7bjwR1ae1WFrIvRy_U0Eic&_nc_zt=14&_nc_ht=scontent-arn2-1.xx&_nc_gid=AjZB38o7I1fVADApQG3ddv_&oh=00_AYAWz-J21PCzXPtZ2RFwzNeMPtko6tmtg3LhS1xaG7Asjw&oe=67CFB37A>; Meta, 'European Union Terrorist Content Online Transparency Report' (Meta 2025) <https://transparency.meta.com/sr/eu-online-report-fb-feb-25>.

[198] TikTok, 'TikTok's EU Terrorist Content Online Regulation (EU) 2021/784 Transparency Report' (TikTok 2025) <https://www.tiktok.com/transparency/en/tco-report-2025>.

[199] TikTok (n 64).

[200] Coche (n 29).

[201] Daphne Keller, 'Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users' (*University of Chicago Law Review Online*, 2022) <https://lawreviewblog.uchicago.edu/2022/06/28/keller-control-over-speech/> accessed 17 July 2024.

[202] 'Gonzalez v. Google: Brief of American Civil Liberties Union, American Civil Liberties Union of Northern California, and Daphne Keller as Amici Curiae in Support of Respondent' (*American Civil Liberties Union*) <https://www.aclu.org/documents/gonzalez-v-google-brief-american-civil-liberties-union-american-civil-liberties-union-northern> accessed 5 June 2025.

[203] Rishabh Kaushal and others, 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database', *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2024) <https://dl.acm.org/doi/10.1145/3630106.3658960> accessed 21 March 2025.

[204] Luca Belli and Jamila Venturini, 'Private Ordering and the Rise of Terms of Service as Cyber-Regulation' (2016) 5 Internet Policy Review <https://policyreview.info/node/441> accessed 5 June 2025.

[205] Palumbo (n 182).

[206] ibid, para 68.

[207] *Ahmet Yildirim v Turkey* (n 81).

[208] *Cumhuriyet Vakfi and Others v Turkey* (n 19).

[209] Case C-284/23 *TC v Firma Haus Jacobus Alten- und Altenpflegeheim gGmbH* [2024] EU:C:2024:558, para 32.

[210] See for instance Case 294/83 *Parti écologiste 'Les Verts' v European Parliament* [1986] EU:C:1986:166; C-222/86 *Unectef v Heylens* [1987] ECR 4097.

[211] *Karastelev and Others v Russia* App no 16435/10 (ECtHR, 6 October 2020), para 118.

in a prescribed manner'. This is coupled with further requirements of independence, impartiality and procedural safeguards.[212]

With that in mind, this section first focuses on explaining the institutional framework from the perspective of independence and impartiality as core elements for the institutional design within Article 6 ECHR and Article 47 of the Charter. Then, this section explains the specific procedural safeguards that are relevant against undue upload filtering.

4.1. The institutional framework

Both Article 6 ECHR and Article 47 of the Charter refer to the right to be heard by an 'independent and impartial tribunal'. Within ECtHR case law, the first of these concepts typically refers to the courts' independence from the other branches of government and from the parties,[213] and is assessed under considerations of 'the manner of appointment of its members and their term of office, the existence of safeguards against extraneous pressure and the question whether the body presents an appearance of independence.'[214] On its side, impartiality refers to 'the absence of prejudice or bias' which, in the context of the right to a fair trial is assessed by determining if the adjudicator had 'any personal prejudice or bias in a given case' or 'hierarchical or other links' between them and other actors within the proceeding.[215] The ECtHR has noted that both independence and impartiality are 'closely linked' and its assessment must be done jointly in some instances.[216] Those notions have been embraced by the CJEU in its case law ;[217] and their application, both by that Court and the ECtHR, has traditionally referred to courts in the traditional sense.

The ECtHR has established that the right to fair trial under Article 6 ECHR provides protections in relation to the institutional organisation of the court system.[218] This should be read in line with the protection given by Article 13 ECHR to an effective remedy, which according to the ECtHR 'guarantees the availability at the national level of a remedy to enforce the substance of the Convention rights and freedoms in whatever form they might happen to be secured in the domestic legal order.'[219] In parallel, the ECtHR has also reasoned that a critical element for safeguarding freedom of expression is that the legal framework provides an adequate 'degree of protection from abuse'.[220]

Nevertheless, as Ortolani has noted, court litigation in the context of content moderation disputes is often inaccessible due to high costs, lengthy proceedings, and legal uncertainties, such as jurisdictional issues, particularly affecting marginalized groups and those with limited resources.[221] In addition to this, content moderation disputes are unprecedented in nature due to factors such as the frequency and scale of user-generated publications that may give rise to conflict.[222] In that sense, a scenario in which the possibility of contesting undue upload filtering is only within the boundaries of judicial redress would create relevant limitations against the rights of users. Furthermore, if there was a policy decision to enclose moderation decisions within courts, States would face the challenge of creating an institutional infrastructure that addresses the particularity of the speed and scale of online discourse by, for instance, creating special courts for moderation with expedited proceedings.

Against that background, the EU legal framework establishes a set of remedial safeguards that can be divided in four layers: The first one relates to internal redress within platforms, the second one relates to alternative dispute resolution (ADR), the third one relates to administrative redress and the last one encompasses judicial redress. In principle, each layer of the multilayer framework applies independently and without prejudice to the others, allowing individuals to seek recourse through any of them without precluding access to the others. However, although the institutional framework does not prevent resorting to courts, the factors mentioned above suggest that most disputes will typically not be resolved at the court level. This is coupled with a clear imbalance of power between platforms and users, as the former are bodies acting with the role of regulators, adjudicators and publishers in the context of disputes about the limits to freedom of expression[223] and are, at the same time, in control of the architecture where speech is disseminated and where decisions on it are enforced.[224]

Having said that, it should be noted that the ECtHR has considered that, in situations where there are limitations to access to courts, there should be 'reasonable alternative means to protect effectively' people's rights.[225] In that sense, the adequate protection of users against undue upload filtering requires not only the ability to resort to courts, but also the proper functioning of the various layers of protection at disposal. Furthermore, as the following paragraphs will show, the second layer of remedial safeguards is the one most likely to gain traction among users and platforms, as it offers an expeditious, user-friendly, and cost-effective procedure. Nevertheless, given its role within a broader institutional framework for rights protection, it must be analyzed in conjunction with the other layers.

4.1.1. The first layer: internal appeals

The first layer of remedial safeguards involves human review and internal appeals.[226] Under Article 20 DSA, users are allowed to appeal any decision against them through a complaint-handling mechanism operating 'under the supervision of appropriately qualified staff, and not solely on the basis of automated means.'[227] If the appeal concerns

[212] *Cyprus v Turkey* App no 25781/94 (ECtHR, 10 May 2001), para 233.
[213] *Beaumartin v France* App no 15287/89 (ECtHR, 24 November 1994), para 38.
[214] *Sacilor Lormines v France* App no 65411/01 (ECtHR, 9 November 2006), para 59.
[215] *Micallef v Malta* App no 17056/06 (ECtHR, 15 October 2009), paras 193-196.
[216] *Kleyn and Others v the Netherlands* Apps no 39343/98, 39651/98, 43147/98 and 46664/99 (ECtHR, 6 May 2003), para 192.
[217] Joined Cases C-585/18, C-624/18 and C-625/18 *A K v Krajowa Rada Sądownictwa, and CP, DO v Sąd Najwyższy* [2019] EU:C:2019:982.
[218] *Golder v the United Kingdom* App no 57496/16 (ECtHR, 21 February 1975), para 32.
[219] *Kaya v Turkey* Apps no 58138/09, 17749/11, 27906/17 (ECtHR, 19 February 1998), para 106.
[220] *OOO Flavus and Others v Russia* Apps no 12468/15 23489/15 19074/16 (ECtHR, 16 November 2020), para 44.
[221] Ortolani (n 39).
[222] Ruairí Harrison, Jonny Shipp and Aebha Curtis, 'The Internet Commission Working Paper: Settling DSA-Related Disputes Outside the Courtroom: The Opportunities and Challenges Presented by Article 21 of the Digital Services Act' (The Internet Commission 2024) <https://tag-craft.files.svdcdn.com/production/assets/assets/Internet-Commission/The-Internet-Commission-Working-Paper-challenges-presented-by-Article-21.pdf> accessed 17 March 2025.
[223] Anni Carlsson, *Constitutional Protection of Freedom of Expression in the Age of Social Media: A Comparative Study* (Department of Law, Uppsala University 2024), pp 119-120.
[224] Tom Divon, Carolina Are and Pam Briggs, 'Platform Gaslighting: A User-Centric Insight into Social Media Corporate Communications of Content Moderation' (2025) 2 Platforms & Society 29768624241303109; Ortolani (n 39).
[225] *Klausecker v Germany* App no 415/07 (ECtHR, 6 January 2015), paras 71-75.
[226] For more details on the incorporation of human intervention in moderation see: Emmanuel Vargas Penagos, 'Platforms on the Hook? EU and Human Rights Requirements for Human Involvement in Content Moderation' (2025) 1 Cambridge Forum on AI: Law and Governance.
[227] DSA, art 20(6).

copyrighted content, it must include human review.[228] This is notwithstanding the obligation to have automated decisions marking content as removable for containing terrorist material subject to human oversight and verification before enforcement.[229]

Despite frequently being referred to as a mechanism of private adjudication, content moderation structures, including their internal appeal procedures, do not qualify as courts or tribunals in the sense of Article 6 ECHR or Article 47 of the EU Charter.[230] The requirements of independence and impartiality are very unlikely to be met at this layer. The appeals mechanisms are typically operated by company-hired personnel or automated tools, which may be either designed, applied, or both, under staff supervision. In that sense, these are private adjudicators that operate with 'instructions in the performance of their judicial duties' by those who designate them, which means that they are not independent.[231] Moreover, moderation structures and policies are created and administered according to the economic and political considerations of their company owners,[232] something that can be seen as at least prone to cause 'extraneous pressure' and affect any appearance of independence. These factors may also indicate the existence of 'hierarchical and other links' between those operating the appeals mechanisms and the upload filter that made the initial decision, which would bar them from being considered impartial.

Regardless, demanding independence and impartiality within platforms' internal moderation structure would run into several practical complications and perhaps impose limitations on the freedom to conduct a business, among other rights. In that sense, the EU legal framework does not ask this layer of the remedial safeguards to be 'independent' or 'impartial', but instead establishes a requirement to act in a 'timely, non-discriminatory, diligent and non-arbitrary manner'.[233] The requirements of 'independence' and 'impartiality' are instead explicitly imposed by the DSA on the second[234] and third[235] layers of remedial safeguards, and by Article 6 ECHR and Article 47 of the Charter on the fourth layer.

4.1.2. *The second layer: alternative dispute resolution*

At the second layer, the EU legal framework establishes the possibility to resort to out-of-court dispute settlement (ODS) as an external appeal mechanism against moderation decisions. Article 21 of the DSA allows this possibility to dispute any moderation decision at any stage and at any body certified by a Digital Services Coordinator of any Member State under certain standards of impartiality and independence, expertise, and operational capacity.[236]

Article 21 DSA complements Article 17(9) of the Copyright Directive, which provided the right to such ODS mechanisms, and which was considered by the CJEU to be one of the safeguards for upload monitoring under that Directive.[237] A similar right to ODS is provided by Article 28b(7) of the Audiovisual Media Services Directive. However, it should be noted that those Directives are limited to mandating the availability of such a redress mechanism, which Member States could just argue as covered with any alternative dispute resolution system they have. Article 21 DSA creates specific obligations for the existence of bodies with dedicated expertise in the content moderation field.

Article 21(2) DSA points out that decisions by ODS bodies are non-binding. Their applicability is limited to the discretion of platforms and merely suggested as a possible measure that VLOPs can implement to address risks identified in their annual mandatory risk-assessments.[238] This was a key modification from the initial draft DSA proposal, which made the decisions by these bodies binding on platforms.[239]

Regardless, ODS can be an attractive alternative to users, particularly because Article 21(5) DSA provides that users have the right to resort to it free of charge or at modest fees that can be reimbursed by the platform if the appealing user is successful in their appeal. Moreover, despite decisions being non-binding, there may be a degree of pressure to comply with them as Article 24(1) mandates platforms to, within their annual reports, include statistics on the number of disputes submitted to these bodies, their outcome and the number of disputes in which the decisions of the body were implemented by the platform.[240] As such, platforms may be compelled to implement decisions out of reputational concerns. In addition to this, platforms may be incentivized to comply to avoid the costs of deploying staff to decide with which decisions from ODS bodies they agree or not. Moderation is a costly procedure, and adding an additional burden would surely be against the platforms' interests.

Articles 21(3)(a) and 21(3)(c) DSA provide some substance to the 'independence' and 'impartiality' requirements of these bodies by establishing that they should be 'financially independent' of platforms and users and not be remunerated in a way that is linked to the outcome of the procedure. However, these criteria may not be enough to assess those requirements.[241] In that sense, Gradoni and Ortolani have recently pointed out that the analysis should not only focus on protecting users from platforms but also platforms from undue influence by their competitors.[242] For instance, they note that the Irish Appeals Centre Europe (ACE) received funds from Meta's Oversight Board and three of its founding directors are trustees of that body. ACE decides on cases from several platforms, which could raise concerns about its impact on competitive dynamics.[243]

Arguably, out-of-court dispute settlement mechanisms provided by the DSA are a form of voluntary arbitration within the ECtHR case law. This Court has considered that only compulsory arbitration falls within the scope of the right to a fair trial under Article 6 ECHR, whereas voluntary arbitration, not mandated by law, does not receive the same protections. However, the ECtHR has also noted that when consent to voluntary arbitration is given in 'take-it-or-leave-it' situations in which individuals must choose between accepting arbitration or being unable to engage in the relevant activity, such arbitration should be regarded as

[228] *Republic of Poland v European Parliament and Council of the European Union* (n 12), para 93; Copyright Directive, art 17(9).
[229] TERREG, art 5(3).
[230] Giulia Gentile, 'Between Online and Offline Due Process: The Digital Services Act' in Annegret Engel, Xavier Groussot and Gunnar Thor Petursson (eds), *New Directions in Digitalisation*, vol 13 (Springer Nature Switzerland 2025) <https://link.springer.com/10.1007/978-3-031-65381-0_11> accessed 22 March 2025.
[231] *Sacilor Lormines v. France* (n 223).
[232] Natali Helberger, 'The Rise of Technology Courts, or: How Technology Companies Re-Invent Adjudication for a Digital World' (2025) 56 Computer Law & Security Review 106118.
[233] DSA, art 20(4).
[234] ibid, art 21(3)(a).
[235] ibid, art 50(1).
[236] DSA, art 21.
[237] *Republic of Poland v European Parliament and Council of the European Union* (n 12), para 93; Copyright Directive, art 17(9).
[238] DSA, art 35, art 36(1)(g).
[239] Jörg Wimmers, 'The Out-of-Court Dispute Settlement Mechanism in the Digital Services Act: A Disservice to Its Own Goals' (2021) 12 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 381.
[240] Ortolani (n 39).
[241] Federica Casarosa, 'Out-of-Court Dispute Settlement Mechanisms for Failures in Content Moderation' (2023) 14 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 391.
[242] Lorenzo Gradoni and Pietro Ortolani, 'Vying for the Scales' (12 March 2025) <https://verfassungsblog.de/vying-for-the-scales-dsa-ace/> accessed 17 March 2025.
[243] These authors illustrated this by examining the Appeals Centre Europe (ACE), designated by the Irish Digital Services Coordinator as an out-of-court dispute resolution mechanism, and which has strong financial and corporate ties to Meta's Oversight Board and Meta itself, while deciding disputes from Facebook, TikTok, and YouTube, with plans to expand to other platforms.

compulsory and must therefore provide fair trial safeguards.[244] These protections, says the ECtHR, are not deprived in cases of 'sui generis' disputes within dispute resolution mechanisms with 'special features'.[245] This is potentially enhanced by the ECtHR's view in *Beg S.P.A. v. Italy*, in which the Court considered that there was no unequivocal waiver of the right to an independent and impartial tribunal in a situation in which a potential conflict of interest of an arbitrator was only evident after the affected party had agreed to the tribunal's jurisdiction.[246] However, as Seatzu and Vargiu point out, the ECtHR, despite expanding the possibilities for fair trial protections within arbitration, did not specify the particular safeguards needed in that context.[247]

Another relevant scenario that should not be discarded is the possibility for platforms and users to agree to 'enforceable settlement' clauses, namely clauses granting alternative dispute resolution a binding nature. The CJEU has considered those clauses to be permissible as long as the users are afforded a real choice and are also granted a right to withdraw from the procedure at any time if dissatisfied with its performance or operation.[248] This is something that may become relevant in this context, because the likelihood of resorting to courts in cases of upload filtering is low due to their various limitations for users. At the same time, there are several factors that turn ODS as a more attractive venue for these disputes. In that sense, although these bodies are in principle not required to have the same level of independence and impartiality as a Court due to their voluntary nature, there may be contextual circumstances that alter the strictness of these requirements.

Another key concern in the context of this layer of redress relates to the applicable law that these bodies are meant to apply. Article 21 DSA does not bring clarity on this and is limited to stating that their purpose is to resolve disputes emerging from moderation decisions. Arguably, Article 14(4) is a basis for allowing these bodies to apply fundamental rights as a legal framework, which would make this redress layer more meaningful for users' protection.[249] The implications and intricacies of these provisions demand further in-depth research.

Finally, the Platform to Business Regulation allows users operating in a professional or commercial capacity in platforms to resort to 'independent' and 'impartial' mediators previously selected by the platform to settle disputes related to moderation.[250] This right was extended to media outlets in specific disputes covered by the European Media Freedom Act.[251]

4.1.3. The third layer: administrative redress

At a third layer are the possibilities to seek redress through administrative means. Article 53 DSA allows users to file complaints before the Digital Services Coordinators of their country of residence or establishment. Those bodies, according to Article 51 DSA, are allowed to impose fines, penalties and to adopt measures to cease infringements (e.g. reinstate content unduly blocked). Furthermore, Article 50(3) provides that, in any case, decisions by Digital Services Coordinators can be subject to judicial review.

The requirements of independence and impartiality at the third layer are established by Article 50(1) and 50(2) DSA, with an emphasis on budgetary independence, the avoidance of both direct and indirect external influence, and a prohibition on receiving instructions from public authorities or private parties. ECtHR case law has established that decisions of administrative bodies that do not comply with the requirements of Article 6 ECHR should be subject to review by judicial authorities with 'full jurisdiction'.[252] The DSA seems to give an avenue to avoid discussions on whether the Digital Services Coordinators fulfil the requirements of Article 6 ECHR by establishing in its Article 50(3) that the conditions for independence set in Article 50(2) 'shall not prevent the exercise of judicial review'. Moreover, this is something that would be aligned with ECtHR case law noting that the 'absence of an effective judicial review may support the finding of a violation' of freedom of expression.[253] This requirement would be stronger in the context of upload filtering as, when related to prior restraints, judicial review is a critical safeguard for preventing abuses of power.[254]

4.1.4. The fourth layer: judicial redress

Judicial redress is not regulated under the EU legislation addressed in this paper, which in any case clarifies the matter by expressly stating that users can access the aforementioned mechanisms without prejudice to their right to go to court.[255] Notably, when reviewing the validity of Article 17 of the Copyright Directive, the CJEU pointed out that the right 'to have access to a court or another relevant judicial authority' was one of the safeguards allowing proportionality for upload filtering under that provision.[256] As can be seen from the preceding paragraphs, the independence and impartiality of the courts responsible for resolving disputes related to upload filtering must meet the highest standards compared to the other layers.

4.2. Procedural safeguards

As has been emphasised, interferences on freedom of expression, including among those prior restraints, must be given with consideration to adequate procedural safeguards. The ECtHR has reasoned that such procedural safeguards apply not only in the context of judicial review of those interferences but also during the initial administrative process in which they are imposed.[257] Arguably, such reasoning is also applicable to moderation decisions in light of the procedural requirements for platforms under the DSA. Moreover, the ECtHR has considered that the lack of procedural fairness in proceedings related to restrictions to freedom of expression can turn into a violation of that right.[258] However, due process requirements for platforms are different from the ones for authorities, particularly because the latter are the ones directly addressed by Articles 6 ECHR and 47 of the Charter.[259]

This subsection focuses on explaining the procedural safeguards that have been traditionally applicable to prior restraints and the way in which they fit into the EU legislative framework related to upload filters. These procedural safeguards become relevant both for the internal design and implementation of filters within moderation systems at platforms, as well as for the assessment done by the remedial layers on

[244] *Mutu and Pechstein v Switzerland* Apps nos 40575/10, 67474/10 (ECtHR, 2 October 2018), paras 113-115.
[245] *Ali Riza and Others v Turkey* Apps nos 30226/10, 5506/16 (ECtHR, 28 January 2020), para 180.
[246] *Beg S.p.a v Italy* App no 5312/11 (ECtHR, 20 May 2021).
[247] Francesco Seatzu and Paolo Vargiu, 'Three Views of a Secret: Missed Opportunities in the Echr's Recent Case-Law on International Commercial Arbitration' (2022) 1 The Italian Review of International and Comparative Law 203.
[248] Case C-75/16 *Livio Menini, Maria Antonia Rampanelli v Banco Popolare Società Cooperativa* [2017] EU:C:2017:457, para 57.
[249] John Albert, 'Expert Insights: Fundamental Rights in DSA Dispute Resolution Procedures - DSA Observatory' (*DSA Observatory*, 20 January 2025) <https://dsa-observatory.eu/2025/01/20/expert-insights-fundamental-rights-in-dsa-dispute-resolution-procedures/> accessed 5 June 2025.
[250] Platform to Business Regulation, art 12.
[251] European Media Freedom Act, art 18(7).
[252] *Ortenberg v Austria* App no 12884/87 (ECtHR, 25 November 1994), para 31.
[253] *Baka v Hungary* App no 20261/12 (ECtHR, 23 June 2016), para 161.
[254] *Kablis v Russia* Apps nos 48310/16, 59663/17 (ECtHR, 30 April 2019) para 92.
[255] See Copyright Directive, art 17(9); TERREG, art 10(2); DSA, art 21(1), art 50(3).
[256] *Republic of Poland v European Parliament and Council of the European Union* (n 12).
[257] *Lombardi Vallauri v Italy* App no 39128/05 (ECtHR, 20 October 2009), para 46.
[258] *Steel and Morris v the United Kingdom* App no 68416/01 (ECtHR, 15 February 2005), para 95.
[259] Gentile (n 240).

appeals filed by persons affected by them.

4.2.1. Reasoning of decisions imposing or upholding upload filters

Article 14(1) DSA establishes an obligation for platforms to include, within their terms of service, information on 'any policies, procedures, measures and tools used for the purpose of content moderation'. In addition to this, Article 17 mandates platforms to provide a statement of reasons for the restrictions imposed to users. Read alongside Article 14 (4) DSA, which establishes that restrictions to users must be applied and enforced 'in a diligent, objective and proportionate manner' and 'with due regard to the rights and legitimate interests of all parties involved', the mentioned obligations can be seen as a mandate to incorporate a human rights standard of providing adequate reasoning to justify any limitations on freedom of expression within the moderation structure.

In that sense, it is well-established ECtHR case law that limitations on freedom of expression must be grounded in 'relevant and sufficient' reasons.[260] In this regard, the CJEU has held that the effectiveness of judicial review requires that restrictive measures on rights be based on reasons that are, in themselves, sufficient to justify the decision.[261] This is crucial to convincingly determine whether the interference complies with the standards for protecting freedom of expression, that is, whether it is 'suitable', 'necessary' and 'proportionate to the legitimate aim pursued'.[262] When referring to the first of these conditions, the CJEU has reasoned that restrictions to rights must be 'appropriate for attaining the legitimate objectives pursued'.[263] Similarly, the ECtHR has declared violations to Article 10 ECHR when the measures 'do not serve to advance' the legitimate aim pursued and instead 'are likely to be counterproductive in achieving' them.[264] The ECtHR has clarified that the second condition 'is not synonymous with "indispensable", neither has it the flexibility of such expressions as "admissible", "ordinary", "useful", "reasonable" or "desirable" and that it implies the existence of a "pressing social need"'.[265] The third of these conditions in the context of the ECtHR case law relates to sufficiently demonstrating that the interference does not undermine the very substance of the right and that no less intrusive alternative measures are available to achieve the same objective.[266] In that vein, restrictions that have a collateral effect of 'rendering large quantities of information inaccessible' on the Internet are considered to be disproportionate.[267]

Against that background, platforms are bound by Articles 14(1), 14 (4) and 17 DSA to establish relevant and sufficient reasons when they decide to voluntarily implement upload filters. Building on this, it can be argued that there is an ostensible pressing social need when the upload filters are mandated by a court injunction that has fulfilled the legal requirements for narrowing down the limit and scope of it, or by a court decision that has declared the illegality of certain content. In the case of upload filters mandated by law, their application should be limited to content that 'manifestly' fits the requirements of that law.

Taking this into account, the use of voluntary filters in cases not mandated by law or court orders cannot be justified by mere considerations of indispensability, admissibility, ordinariness, usefulness, reasonableness, or desirability, and would instead require a determination that no less intrusive alternative measures could achieve the pursued aim of tackling or controlling content that is deemed to be illegal. In that vein, as long as there is no current mandatory obligation to put them in place, voluntary upload filters could be applied in cases of content with potential for causing severe harm (such as child sexual abuse material) in line with some of the platforms' practices summarised in Section II. Besides that, the justification for incorporating voluntary filters for content not deemed illegal but instead violating terms and conditions would demand a higher threshold of suitability, necessity and proportionality. An example of this could be filtering spam, as, despite not being necessarily illegal, would require filtering for allowing a safe, trustworthy and relevant space for platforms' users. Likewise, the use of upload filters to combat content that has been identified as contributing to systemic risks in the context of risk-assessment and mitigation should be limited to situations in which there is a clearly delimited pressing social need for it.

When examining the necessity of restrictions on freedom of expression, the ECtHR has attached particular importance to assessing whether the material in question is of public interest as it would have a stronger protection by Article 10 ECHR.[268] The ECtHR has established that the examination of the necessity and proportionality of restrictions must be done bearing in mind that, although it is possible to impose limitations to journalistic freedom, they must take account of the media's role to 'impart information and ideas on matters of public interest' and not hampering their capacity to discharge their 'public watchdog' role.[269] This is because there is an 'interest of democratic society in ensuring and maintaining a free press'.[270] Such an enhanced protection for media freedom is incorporated by the EU framework for content moderation in the European Media Freedom Act, which establishes specific safeguards in content moderation to media services that have declared to VLOPS that they provide such service.[271] Those protections include specific procedures that are detailed below, but which, as suggested by Recital 50, should also 'duly consider media freedom and media pluralism' in moderation decisions.

Read alongside the CJEU's requirement that filtering mechanisms must distinguish between lawful and unlawful content,[272] and Recital 26 of the DSA, which states that 'good faith' in voluntary measures includes using technology sufficiently reliable to minimize errors, it follows that upload filters deployed by companies must be capable of recognizing when contestable content remains protected for being of public interest. Examples of this could include journalistic reporting on extremism, or publications that may be violent at first sight but relevant for denouncing war crimes. State-of-the-art technologies (e.g. LLMs) have developed some capacities to identify such factors, but they still face limitations in doing it[273] and, as was already mentioned, even the lowest percentage of error can turn into millions of publications subject to an undue limitation.

These are factors that reinforce the need that upload filters are only applied in the most exceptional circumstances that are, in any case, justified as 'relevant and sufficient' reasons under the criteria mentioned in the preceding paragraphs. Furthermore, there would be an emphasised protection against the use of upload filters to screen content by accounts of media services. Following that, if used, upload filters should have the capacity to adequately differentiate contestable and non-contestable content. In parallel, an adequate functioning of the different layers of remedial safeguards would also take such factors into account when deciding to withhold or revoke a restriction based on

[260] *Dilipak v Turkey* App no 29680/05 (ECtHR, 15 September 2015), para 64.
[261] Case T-215/15 *Mykola Yanovych Azarov v Council of the European Union* [2017] EU:T:2017:479, para 136.
[262] *The Sunday Times v the United Kingdom (no. 2)* App no 13166/87 (ECtHR, 26 November 1991), para 50.
[263] Joined Cases C-293/12 and C-594/12 *Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others* [2014] EU:C:2014:238, para 46.
[264] *Bayev and others v Russia* Apps no 67667/09, 44092/12 and 56717/12 (ECtHR, 20 June 2017), para 83.
[265] *The Sunday Times v the United Kingdom (no. 1)* (n 67), para 59.
[266] *Association Rhino and Others v Switzerland* App no 48848/07 (ECtHR, 11 October 2011), para 65.
[267] *OOO Flavus and others v Russia* (n 230), para 43.
[268] *The Sunday Times v the United Kingdom (no. 1)* (n 67), para 65.
[269] *The Sunday Times v the United Kingdom (no. 2)* (n 272), para 50.
[270] *Fressoz and Roire v France* App no 29183/95 (ECtHR, 21 January 1999).
[271] European Media Freedom Act, art 18.
[272] *Scarlet Extended* (n 140); *Netlog* (n 140).
[273] Vargas Penagos, 'ChatGPT, Can You Solve the Content Moderation Dilemma?' (n 45).

upload filtering.

4.2.2. Scope and duration of upload filters

The ECtHR has established that injunctions must provide clarity and certainty about the material they cover.[274] In application of this principle, the ECtHR has considered that establishing temporary bans on printed newspapers that have been declared to have published illegal content is a measure that cannot be considered 'necessary'.[275] In that context, although the ECtHR seems inclined to allow sanctions to dissuade the publication of content similar to that which has been declared to be illegal by a national Court, namely preventing 're-infringements', this can be done through 'less draconian measures' than full bans, 'such as the confiscation of particular issues of the newspapers or restrictions on the publication of specific articles.'[276] The CJEU seems to have ruled at least partly in line with that reasoning when it established that upload filters can be mandated by court injunctions to prevent re-infringements from specific users engaged in that kind of behaviour[277] and in cases of information properly identified as 'equivalent' to material previously declared as illegal.[278]

The ECtHR has considered that, in addition to the obligation to identify content that is deemed illegal, authorities must also provide 'the opportunity to remedy the supposed breach by removing the offending content' to allow an 'informed choice between taking down or modifying the specific content' and object to the order.[279] This may be relevant in the context of upload filters, as platforms could, in cases when they apply filters to screen content 'equivalent' to material declared as illegal, notify the user about that alleged equivalence and allow them to remove the content or contest the decision. A procedure of this sort is applied by X. In some cases, this platform requires users to remove content deemed to be violatory before they can post again, unless they want to appeal. During the appeals process, the contested content is hidden with a label.[280]

Furthermore, the ECtHR has also been cautious when considering the possibility of establishing permanent injunctions. In that sense, interim injunctions should 'not extend beyond a reasonable period'[281] and should be subject to periodic review,[282] and permanent injunctions should be 'subject to review in case of a change of the relevant circumstances'.[283] This can include, for instance, cases in which the passage of time turns the issue into a matter of historical interest.[284] This means that court injunctions mandating upload filters should be limited in time, or, when being of a permanent nature, there should be the possibility of reviewing them with the passage of time.

Likewise, these principles indicate that if platforms choose to implement upload filters as a voluntary measure, they must clearly define and narrow their scope, and their use should, in principle, be limited in time. For example, this could apply during a 'crisis' in accordance with Article 36 DSA. However, it should be noted that it is practically impossible to consider subject matter such as child sexual abuse material or spam as something that should only be targeted temporarily and, as such, the voluntary application of filters could be a permanent measure. Nevertheless, this should not preclude the periodic review of the specific technology applied, taking into account factors such as its effectiveness and error rates.

4.2.3. Transparency of decisions

The ECtHR has established that the lack of transparency of decisions imposing prior restraints constitutes a lack of adequate safeguards against abuse.[285] In this connection, the ECtHR has also established that the non-disclosure of the reasons for a limitation on freedom of expression can result in a restriction of the rights of defence and, therefore, to the right to a fair trial, as it curtails the possibility to substantiate appeals.[286] The requirement in Article 17 DSA for platforms to issue statements of reasons when imposing restrictions on users includes an obligation to inform whether the restriction is based on a violation of the platform's terms of service or on the content being deemed illegal and the specific provision that is argued to have been infringed,[287] the facts and circumstances underlying the decision,[288] whether the decision was made automatically or not,[289] and which are the available avenues for redress.[290] The Platform to Business Regulation establishes separate similar requirements for statements of reasons for users operating in a professional or commercial capacity in platforms.[291]

In addition to this, the European Media Freedom Act requires that, in cases of media service users restricted under a platform's terms of service, the statement must be given *before* the restriction takes effect.[292] Moreover, platforms are required to allow media services to reply to the statement of reasons 'in a meaningful manner' within 24 h, or less in cases of crisis declared on the basis of Article 36 DSA.[293] This would be in line with ECtHR case law establishing that, in cases of injunctions against media publications, a key procedural safeguard for equality of arms would be the ability to submit counter-arguments before they are granted.[294]

In view of the foregoing, the statement of reasons is a key safeguard for ensuring access to the different layers of remedial protection against upload filters. The lack of equality of arms when contesting such measures may render them disproportionate under the ECtHR's case law.[295] In that same sense, disclosure of relevant information about the functioning of the upload filters, such as audits, error rates, among others, would play a key role in the review by external layers of remedial safeguards. This is also critical from an ECtHR perspective, as that Court has considered that the unjustified restriction of access to relevant documentation for substantiating a case can hinder the fairness of proceedings.[296] Such a relevance of the statement of reasons is underpinned by Article 17(4) DSA, which mandates that it must provide information that can 'reasonably allow the recipient of the service concerned to effectively exercise the possibilities for redress'.

However, the previously mentioned analysis by Kaushal et al. found

[274] *Cumhuriyet Vakfi and Others v. Turkey* (n 19), para 62.
[275] *Ürper and others v Turkey* Apps nos 14526/07, 14747/07, 15022/07, 15737/07, 36137/07, 47245/07, 50371/07, 50372/07 and 54637/07 (ECtHR, 20 October 2009), para 44.
[276] ibid, para 43.
[277] *L'Oréal* (n 134)
[278] *Eva Glawischnig-Piesczek* (n 146), para 46.
[279] *OOO Flavus and others v. Russia* (n 230), para 32.
[280] 'When we determine that a post violated the X Rules and the violation is severe enough to warrant post removal, we will require the violator to remove it before they can post again. They will need to go through the process of removing the violating post or appealing our removal request if they believe we made an error. The post will be hidden from public view with a notice during this process.' See: 'Our Range of Enforcement Options for Violations | X Help' <https://help.x.com/en/rules-and-policies/enforcement-options> accessed 21 March 2025.
[281] *Cumhuriyet Vakfi and Others v. Turkey* (n 19), para 66.
[282] ibid, para 64.
[283] *Tierbefreier eV v Germany* [2014] ECtHR 45192/09, para 58.
[284] *Editions Plon v France* [2004] ECtHR 58148/00.
[285] *OOO Flavus and others v. Russia* (n 230), paras 39-43.
[286] *Hadjianastassiou v Greece* App no 12945/87 (ECtHR, 16 December 1992), paras 29-37.
[287] DSA, art 17(3)(d), art 17(3)(e).
[288] DSA, art 17(3)(b).
[289] DSA, art 17(3)(c).
[290] DSA, art 17(3)(f).
[291] Platform to Business Regulation, art 4.
[292] European Media Freedom Act, art 18(4).
[293] ibid, art 18(4)(b).
[294] *Cumhuriyet Vakfi and Others v. Turkey* (n 19), paras 70-74.
[295] *Steel and Morris v the United Kingdom* (n 268), para 95.
[296] *McGinley and Egan v the United Kingdom* Apps nos 21825/93 23414/94 (ECtHR, 9 June 1998), para 86.

that statements of reasons issued by platforms often provide only generic information on the 'decision facts,' with most of them appearing automatically generated and overly brief, preventing users from understanding the grounds for restrictions and limiting their ability to seek redress.[297]

4.2.4. Length of proceedings

The ECtHR has noted that the length of proceedings for reviewing prior restraints should not be 'onerous',[298] particularly because it can substantially undermine the practical effectiveness of judicial review.[299] This turns into a requirement for promptness that can be seen as being covered by the EU legal framework language requiring 'timeliness',[300] 'expeditiousness'[301] or 'reasonableness'[302] in the duration of internal mechanisms for appeal, as well as 'timeliness' for the exercise of functions by Digital Services Coordinators.[303] In the case of ODS, the requirement is more specific, as they are mandated to decide on cases 'within a reasonable period of time and no later than 90 calendar days' and that complex disputes may, at the body's own discretion, be extended for up to an additional 90 days period.[304] Since the ECtHR has established that the 'reasonableness of the length of proceedings is to be assessed in light of the particular circumstances of the case,' ODS procedures should take into account the need to keep the duration of certain procedures related to upload filters as short as possible,[305] for instance in cases of matters of potential public interest or journalistic publications.

The ECtHR has emphasised the need to act more expeditiously in cases of journalistic publications,[306] something that seems embodied in an obligation for platforms to, pursuant to the European Media Freedom Act, take 'all the necessary technical and organisational measures' to ensure that complaints filed by media services within their internal mechanisms are decided 'with priority and without undue delay.'[307] Such priority is not given by the EU framework in other parts of the layers of protection, but should be considered as applicable due to the relevance given by ECtHR case law to act promptly in those cases.

4.2.5. Enforcement of decisions reversing decisions by upload filters

Finally, it should be noted that decisions reversing decisions by upload filters should be enforceable. The ECtHR has emphasised that the right to fair trial entails that judgments should be executable[308] and that refusals, failures or delays by authorities in compliance are against that right.[309] As has been noted so far, decisions by ODS bodies are not binding, which at the same time means that they are not enforceable, different from court or administrative decisions. This would not be the case if an enforceable settlement clause, as explained in sub-Section 4.1, existed. However, since platforms are called under Article 21(2) DSA to participate 'in good faith' in ODS proceedings, they should be expected to take steps necessary to ensure that reversal decisions are enforceable.

Particularly in cases where the restriction applied by an upload filter entails the removal of content, the enforcement of decisions ordering reinstatement would require the preservation of the material by the platform until the specific procedure is over. The ECtHR has noted the importance of allowing the possibility of challenging decisions for the destruction of material before it is effectively destroyed as a relevant safeguard for freedom of expression.[310] Clear rules on this would also be desirable as there have been various instances of users complaining about the lack of clarity on what happens to their content once it is made unavailable.[311]

The DSA does not provide specific rules on this subject matter, except for the fact that it allows Digital Services Coordinators to issue interim measures or request their issuance by national authorities in order to prevent serious harms to users. In addition, the DSA allows users to file a complaint through internal mechanisms within six months, with the possibility of a reversal decision.[312] This suggests that content should be preserved for at least that period and until a final decision is reached. The Regulation remains silent on whether content must be preserved beyond that timeframe. Moreover, although the DSA states that it applies without prejudice to the Regulation on European Production and Preservation Orders for electronic evidence in criminal matters,[313] that Regulation focuses on preserving material for criminal investigations, which is unrelated to users' efforts to seek reinstatement of content through the available layers of protection.[314] In any case, it would be reasonable to consider that part of this issue is covered by national rules for preservation of material subject to court proceedings, which would be available to those resorting to national courts.

TERREG is the only legislation that clearly regulates this issue. This regulation establishes that content removed or disabled must be preserved for six months or for longer periods 'only if and for as long as necessary for ongoing administrative or judicial review proceedings'.[315] In addition to this, the Platform to Business Regulation establishes that, following decisions by internal complaint mechanisms to reinstate content of business users, platforms must provide 'any access to personal or other data, or both, that resulted from its use of the relevant online intermediation services prior to the restriction, suspension or termination having taken effect'.[316]

In that sense, the EU legal framework on content moderation provides some safeguards for preserving material that would afterwards be reinstated, particularly before and as long as an internal appeals or administrative or judicial redress is underway. However, the lack of rules on preservation of material within ODS would cause risks about the effectiveness of those bodies.

5. Conclusions

This article has explained how upload filters, despite being effective for tackling illegal or harmful content, function as digital prior restraints that require strict scrutiny under the safeguards established by the ECtHR. While EU secondary legislation acknowledges the necessity of moderation and even mandates or incentivizes filtering in certain contexts against both legal and illegal content, such measures cannot be divorced from the fundamental rights framework that governs freedom

[297] Kaushal and others (n 213).
[298] *Cumhuriyet Vakfi and Others v. Turkey* (n 19), para 66.
[299] *Association Ekin v France* App no 39288/98 (ECtHR, 17 June 2001).
[300] DSA, art 20(4).
[301] TERREG, art 10(2).
[302] Platform to Business Regulation, art 11(1).
[303] DSA, art 50(1).
[304] DSA, art 21(4).
[305] *Adiletta and Others v Italy* Apps nos 13978/88, 14236/88, 14237/88 (ECtHR, 19 February 1991), para 17.
[306] *Cumhuriyet Vakfi and Others v. Turkey* (n 19), para 73.
[307] European Media Freedom Act, art 18(5).
[308] *Scordino v Italy (no 1)* App no 36813/97 (ECtHR, 29 March 2006), para 196.
[309] *Hornsby v Greece* App no18357/91 (ECtHR, 19 March 1997), para 41.
[310] *Zayidov v Azerbaijan (no 2)* App no ECtHR 5386/10 (ECtHR, 24 March 2022), para 72.
[311] Sarah Myers West, 'Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms' (2018) 20 New Media & Society 4366; Anna Veronica Banchik, 'Disappearing Acts: Content Moderation and Emergent Practices to Preserve at-Risk Human Rights–Related Content' (2021) 23 New Media & Society 1527.
[312] DSA, art 20.
[313] ibid, art 2(4)(i).
[314] Regulation (EU) 2023/1543 of the European Parliament and of the Council of 12 July 2023 on European Production Orders and European Preservation Orders for electronic evidence in criminal proceedings and for the execution of custodial sentences following criminal proceedings [2023] OJ L 191 p. 118-180, art 1.
[315] TERREG, art 6(2).
[316] Platform to Business Regulation, art 4(3).

of expression.

Instead of banning upload filters, the EU legislative framework mitigates the inherent risks of upload filters through a multi-layered system of remedial safeguards, including internal appeals, alternative dispute resolution, administrative redress, and judicial review. However, the mere existence of these layers is not sufficient. Their effectiveness depends on clear procedural safeguards, such as well-reasoned decisions, promptness in handling disputes, transparent communication of decisions with users, and enforceability of decisions reversing undue restrictions. These safeguards should also come with an understanding that, given the risks of the collateral effect of imposing unintended limitations on large amounts of content, upload filters should only be applied in exceptional and limited circumstances.

Although social media platforms enjoy a certain degree of freedom to determine what content is admissible on their services, in today's digital landscape, they increasingly assume quasi-adjudicatory functions, making decisions on matters once reserved to courts, but doing so at unprecedented speed and scale. In this context, the concerns expressed by three U.S. Supreme Court judges over half a century ago regarding the need for careful deliberation in complex cases involving restrictions on speech are more relevant than ever. What is more, the safeguards traditionally required for prior restraints in the analogue world must be adapted to the features of social media communication to avoid setbacks to the protection of freedom of expression.

Funding

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program - Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation. Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work the author used ChatGPT in order to improve the readability and language of the manuscript. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Declaration of competing interest

I have nothing to declare.

Acknowledgements

The author thanks his supervisors, Martin Ebers, Katalin Kelemen and Alberto Giaretta for their support and guidance throughout this research. He also thanks Magnus Strand, Joan Barata Mir, Pietro Ortolani and Silvia Carretta for their feedback on an earlier draft of this article. Additionally, the author would like to thank Andrew Leyden for his help with proofreading, as well as to Raissa Carrillo and Paula Castañeda for their support and encouragement.

Data availability

No data was used for the research described in the article.