Essays

Content Moderation and Freedom of Expression Online

Carolina Perlingieri*

Abstract

This article examines the relationship between freedom of expression by individuals online and content moderation activities carried out by platforms. The study analyses current regulatory frameworks, including that under European law, to assess the limits of compatibility between platform-driven moderation and the rights of platform users. Through an analysis of relevant case law, the article demonstrates the need to ensure aspects of regulatory frameworks that curtail freedom of expression that are consistent with the principles of proportionality and reasonableness.

I. The Legal and Technological Regulation of the Relationship Between Users and Platform Operators

A discussion of platforms and 'freedom of expression online' raises numerous questions regarding the close relationship between the medium through which content is disseminated and the content itself. One key issue is the emphasis that should be placed on content moderation[1] by platforms, specifically addressing the limits of its compatibility with the right to freedom of expression.

To adequately address this topic, it is essential to consider the legal and technological regulation of the user-platform relationship, enriched by contributions from European Union law, particularly under the Digital Services Act (DSA),[2] which is legislation specifically designed to apply to the technological infrastructure of

* Full Professor of Private Law and Law of New Technologies, University of Naples Federico II.

[1] On the content moderation activities of digital platforms, see T. Gillespie, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media* (New Haven: Yale University Press, 2018), 5; Id, 'Content moderation, AI, and the question of scale' available at https://tinyurl.com/y99kscpa (last visited 31 January 2026). On the tools used to achieve content moderation, see J. Grimmelmann, 'The Virtues of Moderation' 17 *Yale Journal of Law and Technology*, 61 (2015).

[2] European Parliament and Council Regulation (EU) 2022/2065. For a review of the different stages in the development of the Regulation see S.F. Schwemer, 'Digital Services Act: A Reform of the e-Commerce Directive and Much More' *Research Handbook on EU Internet Law*, 1 (2022); O. Pollicino, 'Verso il Digital Services Act: problemi e prospettive. Presentazione del simposio', available at https://tinyurl.com/ybuzzwfs (last visited 31 January 2026); A. Nicita, 'Le piattaforme online tra moderazione e autoregolazione: verso il Digital Services Act', available at https://tinyurl.com/zexfjyc6 (last visited 31 January 2026). On the distribution of responsibilities to platforms by the Digital services Act see P. Stanzione, 'Introduzione', in Id ed, *I "poteri privati" delle piattaforme e le nuove frontiere della privacy* (Torino: Giappichelli, 2022), 9.

digital services.

The relationship between users and platforms must be examined starting from the moment that access to the platform by the user is granted by the platform operator, based on a contract formed through consent expressed by the user upon registration for a service.[3] This act constitutes acceptance of a contractual offer, ie, under Italian law a public offer (under Art 1336 of the Italian Civil Code) contained in General Terms of Use between the user and platform.

Scholarly reflections on this first focused on analysing the contents of the DSA and the standards set by Big Tech companies, which continue to evolve over time. This inquiry then expanded to scrutiny of the technical rules – also defined unilaterally[4] – governing various online activities (social, political, economic, religious, or recreational).

Platforms not only influence the formation of the user-platform relationship but also its development in both legal and technological terms, as platforms retain the authority to set legal and technical rules, implemented through operational algorithms that govern functioning of the platform, as well as through technological tools that enforce both legal rules and community standards. For example, some tools block access to illicit content, preventing access and sharing in order to enforce content control (from the point of view of legality and appropriateness), while others disable access as a means of interrupting the service.

Content moderation must therefore be examined through the lens of the contractual nature of the user-platform relationship.[5] This perspective frames content moderation as an exercise of private authority[6] governed by contractual terms, including self-regulation and co-regulation. Consequently, controls must be applied concerning the legality, abusiveness, and appropriateness of content, both within the General Terms of Use and as a result of the operational dynamics governed by rules of the platform, including those set by algorithms.

[3] On the consent given at the time of registration to the social platform as a manifestation of will and therefore acceptance of a contractual proposal of the operator and not as consent to processing, see C. Perlingieri, *Social Networks and Private Law* (Napoli: Edizioni Scientifiche Italiane, 2017), 64.

[4] ibid 23; F. Resta, 'Le piattaforme e la visibilità del potere', in P. Stanzione ed, *I "poteri privati"* n 2 above, 369; A. Simoncini, 'La co-regolamentazione delle piattaforme digitali' *Rivista trimestrale di diritto pubblico*, 1031 (2022).

[5] C. Perlingieri, *Social Networks* n 3 above, 85; C. Pinelli and U. Ruffolo, *I diritti nelle piattaforme* (Torino: Giappichelli, 2023), 46; G. Alpa, 'Sul potere contrattuale delle piattaforme digitali' *Contratto e impresa*, 721 (2022).

[6] G. Teubner, 'Regimi privati globali. Nuovo diritto spontaneo e costituzione duale nelle sfere autonome della società globale', in Id, *La cultura del diritto nell'epoca della globalizzazione. L'emergere delle costituzioni civili* (Roma: Armando, 2005), 59; Id, *Costituzioni societarie: politica e diritto oltre lo Stato* (Milano: Franco Angeli, 2011); Id, *Nuovi conflitti costituzionali. Norme fondamentali dei regimi transnazionali* (Milano: Mondadori, 2012).

II. The Contribution of European Legislation to the Proceduralisation of Content Moderation

An important contribution to this sphere comes from the DSA,[7] which introduced a system for proceduralising content moderation activities by laying down limits and rules that platforms must follow in their content moderation practices.

The rationale for a differentiated regime specifically targeting Big Tech companies lies in the power they exert over digital communication regulation.[8] This necessitates scaling back their influence to ensure due cooperation with public authorities through models of private governance, as exemplified by the due diligence rules (Arts 11-48 DSA).

The following are the key intervention areas for all platforms under the DSA:

1. Harmonised procedures for monitoring illegal content, to ensure diligent platform behaviour (Arts 7-10 DSA);

2. Transparency and accountability obligations regarding content removal, designed to involve users by providing them with detailed explanations for decisions of providers (Arts 15-17 DSA);

3. Additional obligations for very large online platforms and very large online search engines to manage 'systemic risks' arising from service design and usage, ensuring intended by-design functioning (Arts 34-37 DSA and Recital 79);

4. Lastly, the creation of new independent authorities, termed 'Coordinators of Digital Services,' vested with broad oversight powers to monitor and evaluate compliance with the DSA (Art 51 DSA). This aspect reflects a remedial strategy that extends beyond the judicial authority's provision of remedies (mainly through the recognition of the invalidity of contracts) to include administrative oversight, with various types of actions, including the imposition of sanctions.

This regulatory framework appears commendable, as it targets the procedural and organisational mechanisms of platforms that require cooperation. However, this cooperation might not always be fully realised. For instance, transparency obligations should extend not only to access by regulators to platforms' databases but also to their algorithm systems. Notably, in 2018, the European Commission

[7] See the recent article R. Razzante, 'Nuove frontiere della libertà d'espressione alla luce del Digital Services Act (DSA) e dell'evoluzione normativa europea, tra criticità applicative e possibili risvolti costituzionali', available at https://tinyurl.com/5f3dx3bn (last visited 31 January 2026); A. Palumbo and J. Piemonte, 'Delega di funzioni regolamentari e lotta ai rischi sistemici causati della disinformazione nel Digital Services Act: quali rischi per la libertà di espressione?' available at https://tinyurl.com/3juhafba (last visited 31 January 2026); M. Astone, 'Digital Services Act e nuovo quadro di esenzione della responsabilità dei prestatori di servizi intermediari: quali prospettive?' *Contratto e impresa*, 1050 (2022).

[8] See K. Klonick, 'The new governor: the people, rules, and processes governing online speech' 131 *Harward Law Review* 1662 (2018); G. Resta, 'Diritti fondamentali e diritto privato nel contesto digitale', in Id and F. Caggia ed, *I diritti fondamentali in Europa e il diritto privato*, available at https://tinyurl.com/fytp27kp (last visited 31 January 2026), 117; J.M. Balkin, 'Free Speech is a Triangle' 118 *Columbia Law Review*, 2011 (2018); Id, 'How To Regulate (And Not Regulate) Social Media' 1 *Journal of Free Speech Law*, 71 (2021).

adopted the Code of Practice on Disinformation,[9] which was voluntarily adopted by Facebook, Google, Twitter, and Mozilla. However, these companies refused to share strategic information that would have enabled researchers to study the functioning of their algorithms, subsequently reaffirming this refusal during revision of the Code in 2022 when it became a code of conduct under the DSA.[10] These platforms merely provided 'facilitated access to data for researchers'.

While this 'facilitated' access to datasets is undoubtedly a first step towards understanding the functioning of the technological rules of platforms, it remains insufficient unless it is coupled with facilitated access to the operational criteria of the algorithms, particularly hyperparameters, which identify and select the data inputs necessary to achieve the desired outcomes, and which is a critical need for those training the systems.

III. The Impact of the AI Act on the Operation of Algorithmic Systems on Platforms

The issue raised prompts a further question: What impact might the Artificial Intelligence Act[11] (AI Act) have on the functioning of platforms' algorithmic systems? Can its principles[12] be invoked to enforce the transparency obligations introduced by the DSA thereby ensuring its reliability, including for very large online platforms and search engines? The DSA had already introduced additional obligations to manage 'systemic risks' and verify the intended by-design functionalities for such platforms and search engines.

These risks, which may bring about adverse effects, are grouped into the following four areas: a) the exercise of fundamental rights; b) civic discourse, electoral processes, and public safety; c) gender-based violence, public health and child protection; and d) severe harm to individuals' physical and mental well-being.

These systemic risks mirror the high-risk systems regulated by the AI Act, which imposes transparency requirements in several areas concerning: a) the original purpose of AI data collection, the use of algorithms, and the system's intended users; b) datasets, to ensure 'statistical quality'; c) synthetic content, with labelling requirements to indicate that output has been created or manipulated artificially; and, especially, d) the functioning of these systems, where input legality depends on system traceability throughout its lifecycle via 'automatic event logging'.

It therefore appears possible to extend these transparency obligations under

[9] The code is available at https://tinyurl.com/mpd9k2zt.

[10] https://tinyurl.com/ypd944d2.

[11] European Parliament and Council Regulation (EU) 2024/1689 (Artificial Intelligence Act) laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, available at eur-lex.europa.eu.

[12] See the recent article on this topic C. Perlingieri, 'Intelligenza artificiale tra princípi e regole', in Ead, *Innovazione tecnologica e diritto civile. Saggi* (Napoli: Edizioni Scientifiche Italiane, 2025), 187.

the AI Act to the 'systemic risks' under the DSA that arise from the design and use of large platform services. as This is because the transparency obligations can be equated with systemic risks, and because both are expressions of the principle of reliability. Accordingly, very large online platforms and search engines should also be subjected to transparency obligations, particularly concerning the functioning of algorithms, ensuring access to the operational criteria used to select data for achieving desired outcomes.

Such transparency obligations, grounded in the principle of reliability, allow for scrutiny not only of the General Terms of Use but also of algorithmic regulatory procedures, the technical functions of the platforms, and the outcomes of regulatory activities. This scrutiny becomes critical when algorithmic processes determine the search and filtering of unlawful content, operating on the basis of predefined criteria (eg, the name of the individual involved, the circumstances in which the violation was identified, or the identification of content identical or equivalent to that declared unlawful).

The automatic nature of filtering algorithms may prove inadequate, at times producing incomplete results by failing to identify unlawful content, and at other times, erroneous responses, mistakenly removing lawful content. Ensuring transparency of algorithmic operations would allow the jurist to interpret the outcome of the system or, otherwise, decide not to use it.

In conclusion, a new collaborative role for providers emerges, as they must ensure compliance with contractual obligations. This simultaneously requires scrutiny of content moderation, from both legal and technological points of view in respect of their relations with users.

IV. Practical Implications of the Collaborative Role of Providers in Content Moderation

We can now examine the practical implications of this approach, particularly regarding content moderation, which affects freedom of expression.[13] Concerning the fulfilment of contractual obligations, scrutiny of content moderation must take into account the regulatory framework which governs the platform's economic activity.[14] Therefore, first and foremost, the constitutional norms of the user's jurisdiction must be considered, meaning that the controls will have to ensure

[13] On this topic, see the recent article J. Hawkins, *Content Moderation Issues* (Publifye AS, 2025), exploring the challenges of managing online content in the digital age.

[14] C. Perlingieri, 'Libertà di espressione e di informazione nella comunicazione digitale', in Ead, *Innovazione tecnologica* n 12 above, 115. On this topic, see also F. Marongiu Buonaiuti, 'L'ambito di applicazione territoriale del Digital Markets Act e del Digital Services Act. Tra applicazione "extraterritoriale" del diritto dell'Unione europea e attribuzione di un carattere internazionalmente imperativo alle sue norme', in G. Caggiano et al eds, *Verso una legislazione europea su mercati e servizi digitali* (Bari: Cacucci, 2021), 171, who highlights the 'necessary application' of the regulation of digital services markets regardless of the governing law.

three things: legality, fairness, and justifiability. One cannot, for example, agree with Facebook's position in the case of the shutdown of a political movement's page on the social media. The platform argued that its actions were not bound by the constitutional values of the state in which it operates, claiming that it was completely irrelevant to the deactivation of an association's pages that it was an organization prohibited by Italian law. This has the further consequence that not even an express provision of law could limit its right to decide which content to host and which to exclude, stating that unilateral general conditions of use prevail in absolute terms.

The imposition of a penalty that undermines freedom of expression cannot be invoked solely on the basis of the breach of community standards regarding contracts considered in their entirety and without regard to external legal or constitutional constraints.

For instance, posts inciting hate, violence - including that owing to racism - xenophobia, and discrimination against minorities, praising Fascism, Nazism, or their symbols, are unlawful not because they violate contractual clauses, but because they conflict with the constitutional legal system of a State where the operator addresses its activities. A conflict of this kind can justify taking down a social media page, even if members of the movement use it as a means of exercising their freedom of political expression.[15]

V. Observations and Conclusions in the Light of the European Court of Human Rights' Ruling on Hate Speech

The perspective of dual scrutiny of both legal and technical regulation proves to be particularly useful when reflecting on an incident involving hate speech.

The incident involved two young people who publicly burned a photograph of members of the Spanish royal family to contest and express strong political criticism towards the Spanish monarchy. The Spanish Constitutional Court initially rejected the accused's appeal, deeming the act a form of incitement of hatred and violence against the King and the monarchy. The two individuals were subsequently convicted by Spanish courts for insulting the Crown, prompting them to seek intervention by the European Court of Human Rights, claiming an infringement

[15] See the 'Casapound case': the Court of Rome, in its decision of 3 December 2022, aligned with the position of Forza Nuova, revising the decision made in the pretrial phase, where the Court had highlighted that 'the determination of the violation of the fundamental principles of association and freedom of expression should be made through a full review process'. On this topic, see C. Perlingieri, 'Libertà' n 13 above, 116. In case law, see the recent decision by the Tribunale di Roma 5 December 2022 no 17909, available at https://tinyurl.com/mtyyza3b (last visited 31 January 2026), with note by A. Golia, 'La sentenza del Tribunale di Roma sulla rimozione dei profili social di Casapound da parte di Meta', available at https://tinyurl.com/362kmuk8 (last visited 31 January 2026); G.E. Vigevani, 'Dal "caso Casapound" del 2019 alla "sentenza Casapound" del 2022: piattaforme digitali, libertà d'espressione e odio on line nella giurisprudenza italiana', available at https://tinyurl.com/ta387tey (last visited 31 January 2026).

of their right to political expression.

Particularly interesting is the decision of the European Court,[16] which took the opportunity to distinguish between non-violent acts related to political protest and acts of incitement of hatred or violence. In its ruling, the Court emphasised that restrictions on freedom of expression should be interpreted restrictively, especially with regard to political debate.[17] Therefore, considering that Art 10 European Convention on Human Rights (ECHR) protects not only inoffensive expressions but also those that disturb, offend, or shock, the act in question expressed an idea linked to a public debate on the independence of Catalonia and thus contributed to ensuring pluralism.

I consider that the principles enunciated in this case could also be useful for evaluating dissemination of information online. The question is: Would the content filtering and search algorithm have flagged the relevant content, for example, an image or video showing the two individuals burning the royal photograph, as an act of violence or incitement to hatred? Would the social platform have removed the content or closed the individuals' pages? In conclusion, could the algorithm have assigned a meaning to the act, distinguishing between a non-violent act of political protest and an act of incitement of hatred or violence? I think that the issue of content moderation and its limitations must also address the outcomes produced by algorithms governing platform operations, as algorithmic decisions rely on syntactical analysis rather than semantic analysis. This means that the algorithm is incapable of grasping the variety of possible meanings and is unable to ensure compliance with constitutional legality.[18]

In conclusion, we must continue in the direction initiated by European lawmakers. It is essential to identify measures that are appropriate for the current technological context while also ensuring remedies that adhere to the principles of proportionality and reasonableness.[19] This can only be guaranteed by an interpreter of the law sensitive to functional and axiological considerations.

[16] European Court of Human Rights, *Stern Taulats and Roura Capellera* v *Spain* App nos 51168/2015 and 51186/2015, Judgment of 13 March 2018, available at hudoc.echr.coe.int.

[17] For an overview of the pronouncements of the European Court of Human Rights that have marked the evolutionary path on the subject, see R. Petruso, 'Responsabilità delle piattaforme online, oscuramento di siti web e libertà di espressione nella giurisprudenza della Corte Europea dei Diritti dell'Uomo' *Diritto dell'informatica e dell'informazione*, 511 (2018); G. Gorrias Lucente, 'Internet e libertà di manifestazione del pensiero' *Diritto dell'informatica e dell'informazione*, 597 (2000); C. Melzi D'eril, 'La complessa individuazione dei limiti alla manifestazione del pensiero in internet' *Diritto dell'informatica e dell'informazione*, 571 (2011); E. Laidlaw, *Regulating Speech in Cyberspace* (Cambridge: Cambridge University Press, 2015), 46.

[18] On the principle of constitutional legality and the difference between 'legality' and 'legitimacy' see P. Perlingieri, *Il diritto civile nella legalità costituzionale secondo il sistema italo-europeo delle fonti*, II, *Fonti e interpretazione* (Napoli: Edizioni Scientifiche Italiane, 4th ed, 2020), 191. Otherwise, for an equivalence between legality and legitimacy see N. Irti, *Società civile. Elementi per un'analisi di diritto privato* (Giuffré: Milano, 1992), 166; Id, 'Quattro giuristi del nostro tempo' *Rivista di diritto privato*, 768 (1998).

[19] On this topic, see G. Perlingieri, *Profili applicativi della ragionevolezza nel diritto civile* (Edizioni Scientifiche Italiane: Napoli, 2015), 132.