GREEN PAPER

ON

THE DIGITAL SERVICES ACT AND THE NEED FOR DUE PROCESS IN ALGORITHMIC CONTENT MODERATION

To: Mr. Roberto Viola, Director-General of DG Connect (European Commission)

From: A N Other

Date: 31st August 2021

Executive Summary

This green paper aims at discussing how the Digital Services Act proposal is a significant breakthrough in how online content moderation is currently framed at the EU level. The growing use of automated tools, together with a legal framework that pressures intermediary providers to over-block illegal content to avoid liability, has exacerbated the long-standing inadequacies of algorithmic content moderation, namely its lack of accuracy and due process. Indeed, without proper transparency, contestability and accountability, automated moderation has become a tool of private censorship and profit-driven speech regulation, by which users' fundamental rights fail to be safeguarded. Based on these premises, the Digital Services Act, by upholding the liability regime of the E-Commerce Directive and finally introducing tailored mandatory due process obligations for online platforms, can finally achieve a European digital space underpinned by the respect of users' freedom of expression. After an outline of the drawbacks of online content moderation, notably the automated type, and an analysis of the flaws of the current legal framework, this paper discusses how the Digital Services Act proposal aims at achieving a fundamental rights-compliant algorithmic content moderation. Based on such analysis, I suggest several amendments to overcome its existing shortcomings, focusing on restricting the circumstances in which intermediary providers' liability can be triggered, and on strengthening algorithmic transparency and accountability.

Table of Contents

Introduction

Online platforms,[1] especially social media, have grown so much in size and influence recently that they have become actual public spaces,[2] where individuals express their opinions and share content and ideas.[3] Although reluctantly, platforms eventually have had to deal with the issue of content moderation: as Gillespie puts it, 'having in many ways taken custody of the web, [platforms] now find themselves its custodians'.[4]

The difficulty with content moderation is that, if too loose, it results in under-blocking of illegal content, and if too strict, it leads to over-blocking of user-generated content.[5] Moreover, the amount of content circulating online and the pressure from governments to remove illegal content proactively and expeditiously to avoid liability, resulted in the sacrifice of accuracy over speed[6] and led content moderation to lean on the over-blocking side of the spectrum.

While over-blocking alone is capable of jeopardising users' fundamental rights, the emergence of mostly unregulated automated means of moderation has put such fundamental rights under even greater pressure. Notably, algorithmic moderation is often either inaccurate or discriminatory, and human moderators are too few – if not used at all – to correct such mistakes.[7]

Furthermore, the lack of due process in automated moderation, exacerbated by its lack of transparency, prevents users from effectively challenging platforms' decisions.[8] This, in turn, makes

[1] The terminology used through this paper is meant to reflect that of the Digital Services Act, whose art 2(h) defines online platforms as providers of hosting services which store and disseminate to the public information. See European Commission, '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' COM (2020) 825 final ('DSA'). Due to their size and influence, this paper will mainly discuss online platforms, although reference to intermediary providers in general will often be made, especially in terms of liability and its relation to over-blocking.

[2] Alexandre De Streel and others, *Online Platforms' Moderation of Illegal Content Online* (European Parliament 2020) 81.

[3] Giancarlo Frosio and Christophe Geiger, 'Taking Fundamental Rights Seriously in the DSA's Platform Liability Regime' (2020) European Law Journal (forthcoming) 6 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3747756> accessed 20 August 2021.

[4] Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media* (Yale University Press 2018) 5.

[5] Frederick Mostert, ''Digital due process': a need for online justice' (2020) 15(5) Journal of Intellectual Property Law & Practice 378, 378.

[6] Hannah Bloch-Wehba, 'Global Platform Governance: Private Power in the Shadow of the State' (2019)72 Smu L. Rev. 27, 78.

[7] Yifat Nahmias and Maayan Perel, 'The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations' (2021) 58 Harv J on Legis 145, 171-173.

[8] Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011, 2025.

content moderation unaccountable, transforming platforms into 'prosecutor[s], judge[s], jur[ies], and executioner[s]'.[9]

While international organisations and scholars have long advocated for due process standards to be included in platform moderation, the EU and national governments have pursued a political strategy focused on imposing or encouraging platforms to proactively moderate illegal content, leaving due process to self-regulation, out of legislative reach. Ultimately, private censorship has been institutionalised under government control.[10]

For these reasons, the time is ripe for an overarching regulation on content moderation inspired by 'digital due process' principles.[11] Therefore, I argue that the recent DSA proposal, with its horizontal, asymmetric obligations on platforms ensuring a transparent, contestable, and accountable content moderation, finally puts due process and fundamental rights at the core of the political debate. However, a few points must still be addressed to enable the DSA to achieve its full potential.

In this paper, I discuss the reasons why the current content moderation framework – notably the algorithmic one – fails at guaranteeing due process and users' fundamental rights, and thus why the DSA – with some refinements – is a landmark step in Internet speech regulation.

Chapter 1 provides an overview of the different forms of content moderation and the advantages and disadvantages of automated moderation, arguing how its inaccuracy, bias, and lack of due process encroach upon users' fundamental rights and constitute a form of private censorship.

Chapter 2 discusses how despite the calls from international organisations and scholars, recent EU and national initiatives failed to uphold due process, left to ineffective self-regulatory measures. By prioritising the swift removal of illegal content, private censorship has been ultimately institutionalised.

Chapter 3 analyses the DSA proposal, breaking down its provisions into scope, liability, and due process obligations, and suggests amendments to further safeguard due process and users' fundamental rights, for instance by restricting the circumstances in which intermediary providers' liability can be triggered and by strengthening algorithmic transparency and accountability.

[9] Mostert (n5) 385.
[10] Molly K Land, 'Regulating Private Harms Online: Content Regulation under Human Rights Law' in Sandra Braman (ed), *Human Rights in the Age of Platforms* (MIT Press 2019) 294.
[11] Mostert (n5) 379.

1. Online content moderation

Moderation is a broad term that encompasses both 'hard' and 'soft' moderation.[12] While the former addresses the removal of infringing content, that is either illegal, harmful,[13] or contrary to platforms' ToS, the latter includes the way content is organised and offered to users by recommender systems. This paper will address the first type of moderation, with a particular focus on the one performed by algorithms.

1.1 Reactive and proactive moderation

Platforms have traditionally employed a *reactive* content moderation,[14] built on a notice-and-take-down procedure prompted by users' flagging, which aims at removing the infringing content[15] and ultimately at setting the rules of community engagement.[16] Such conduct has initially been endorsed by governments, who chose a 'hands-off approach' to online speech regulation,[17] exemplified by the E-Commerce Directive,[18] whereby platforms are exempt from liability if they *reactively* remove illegal content after having obtained knowledge thereof.[19]

However, due to growing pressure from politics, especially in response to shocking incidents like the terrorist attack that took place in Christchurch (New Zealand) in 2019 and was live-streamed on Facebook,[20] platforms started to rely on a *proactive* approach, whereby they do not (just) wait for users' flagging, but they proactively moderate content.[21] For similar reasons, proactive moderation has progressively been subject to legal obligations.[22] Examples include the controversial German

[12] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7(1) Big Data & Society 1, 3 <https://journals.sagepub.com/doi/full/10.1177/2053951719897945> accessed 20 August 2021.
[13] Illegal content refers to content considered illegal by law, whereas harmful content, while not illegal *per se*, is capable of offending other people.
[14] Cambridge Consultants, 'Use of AI in online content moderation - Report produced on behalf of OFCOM' (*OFCOM*, 18 July 2019) 35 <https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf> accessed 20 August 2021.
[15] Jennifer Cobbe, 'Algorithmic Censorship by Social Platforms: Power and Resistance' (2020) Philos. Technol 2 <https://doi.org/10.1007/s13347-020-00429-0> accessed 20 August 2021.
[16] Gorwa (n12) 3.
[17] Mostert (n5) 379.
[18] 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 ('ECD') art 14.
[19] Giovanni Sartor and Andrea Loreggia, *The impact of algorithms for content filtering or moderation - "Upload filters"* (European Parliament 2020) 26.
[20] Gorwa (n12) 1.
[21] Cambridge Consultants (n14) 35-36.
[22] Sartor (n19) 32.

Network Enforcement Act[23] and the French Avia Law,[24] albeit the latter was shortly after declared unconstitutional.[25]

Proactive moderation can be distinguished into *ex-ante* or *ex-post*, depending on the moment in which the content is removed. The former consists indeed in checking the content before its upload, usually through algorithms.[26] Being akin to censorship, I submit that this type of moderation is the most dangerous for users' fundamental rights, and it should never be imposed by regulators.

However, it is not always straightforward to distinguish between these two types of moderation in the legislation. In the EU, for instance, while the ECD, the baseline regime on illegal content moderation, is built upon a notice-and-takedown mechanism and the prohibition of general monitoring obligations,[27] subsequent sectoral legislation – notably the Audiovisual Media Services Directive[28] and the infamous Directive on copyright in the Digital Single Market[29] – is more ambiguous. Indeed, despite being both without prejudice to Article 15 ECD, the wording of their Articles 28b(1)[30] and 17(4),[31] respectively, seem to require a level of proactiveness much closer to *ex-ante* moderation.[32]

1.2 Algorithmic proactive moderation

This type of legislation does not oblige platforms to choose a particular means of moderation over another. However, pressure from governments and users to have a safe online space[33] and the sheer amount of content circulating online[34] has led platforms to increasingly rely on automatic tools to

[23] Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act) ('NetzDG').
[24] *Loi* n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet.
[25] Sénat, 'Censure de la loi AVIA: il faut combattre la haine sur internet sans fragiliser la liberté d'expression' (Sénat, 19 June 2020) <https://www.senat.fr/presse/cp20200619b.html> accessed 20 August 2021.
[26] Cambridge Consultants (n14) 35-36.
[27] ECD, arts 11-15.
[28] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services [2018] OJ L 303 ('AVMSD').
[29] 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 ('Copyright Directive').
[30] Art. 28b(1) AVMSD requires video-sharing platforms to 'protect' minors and the general public from certain types of content.
[31] Art. 17(4) Copyright Directive requires providers to 'make best efforts to ensure the unavailability' of copyright infringing content.
[32] Sartor (n19) 59.
[33] Emma J Llansó, 'No amount of "AI" in content moderation will solve filtering's prior-restraint problem' (2020) 7(1) Big Data & Society 1, 2 <https://journals.sagepub.com/doi/full/10.1177/2053951720920686> accessed 20 August 2021.
[34] Cambridge Consultants (n14) 38.

moderate content.[35] Moreover, this trend has been further exacerbated during the COVID-19 pandemic.[36]

Automated content moderation includes the use of several tools, from simpler algorithms to complex machine learning models. Common technologies range from hash-based filters, useful to swiftly remove identical 'matched' content, to classifying systems, capable of labelling content, such as hate speech and nudity, and generative adversarial networks (GANs), used to detect manipulated images or videos.[37]

1.2.1 Advantages

Automated content moderation offers some evident advantages compared to that conducted by human operators. To begin with, the outstanding scale of content uploaded on platforms cannot be tackled solely by human moderators.[38] Furthermore, algorithms are highly effective in dealing with live content, where swift detection is paramount, and in removing identical content, avoiding the 'Whack-a-Mole' effect.[39] In addition, algorithmic moderation helps to reduce the psychological damage human moderators are exposed to when assessing the most disturbing content.[40]

These factors explain why proactive automated moderation is now clearly favoured by platforms over traditional users' flagging: for instance, between January and March 2021, around 95% of videos removed by YouTube – ie 9,091,315 – have been flagged by automatic tools.[41] Similar figures can be observed in Facebook's report.[42]

[35] Emma Llansó, Joris van Hoboken and Jaron Harambam, 'Artificial Intelligence, Content Moderation, and Freedom of Expression' (2020) Transatlantic Working Group on Content Moderation Online and Freedom of Expression 3 <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 20 August 2021.
[36] De Streel (n2) 64.
[37] See in general Sartor (n19) 35-44. See also Cambridge Consultants (n14) 49, Llansó and others (n35) 5-7.
[38] Tarleton Gillespie, 'Content moderation, AI, and the question of scale' (2020) 7(2) Big Data & Society 1, 2 <https://journals.sagepub.com/doi/full/10.1177/2053951720943234> accessed 20 August 2021.
[39] Mostert (n5) 381.
[40] Cambridge Consultants (n14) 43-44.
[41] Google, 'YouTube Community Guidelines enforcement' (*Google*, 2021) <https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB> accessed 20 August 2021.
[42] Facebook, 'Community Standards Enforcement Report' (*Facebook*, 2021) <https://transparency.fb.com/data/community-standards-enforcement> accessed 20 August 2021.

1.2.2 Disadvantages

The abovementioned features must not, however, be overstated,[43] if only because automated content moderation not only 'make[s] many, many mistakes',[44] with several 'false positives' or 'false negatives',[45] but also encroaches upon users' fundamental rights in several ways.

To begin with, moderation usually requires a contextual and culturally sensitive assessment.[46] Algorithms lack such context awareness and have been observed to struggle, for instance, with hate speech[47] and memes.[48] Moreover, AI tools are vulnerable to malicious attacks, among which 'adversarial attacks' with GANs have proven to be capable of deceiving even the most advanced automated filters.[49]

This lack of accuracy and reliability is responsible for the simultaneous over-blocking of lawful content and under-blocking of illegal content.[50] Furthermore, automated moderation can be as biased as its human counterpart,[51] leading to discrimination.[52]

Finally, the opaque nature of algorithms makes mistakes and bias even more difficult to detect. This leads to what I submit to be the main drawback of automated moderation, which is its lack of transparency. If content moderation, in general, is quite secretive, that run by algorithms – constituting themselves a 'black box'[53] – can even be inscrutable.[54]

Indeed, the opacity of algorithms makes it almost impossible for users to understand the reasons behind automated takedowns.[55] As a result, the chance of successfully lodging a complaint becomes

[43] Cobbe (n15) 3.
[44] Svea Windwehr and Christoph Schmon, 'Our EU Policy Principles: Procedural Justice' (*EFF*, 27 July 2020) <https://www.eff.org/deeplinks/2020/07/our-eu-policy-principles-procedural-justice> accessed 20 August 2021.
[45] Cambridge Consultants (n14) 37.
[46] Cobbe (n15) 2, Llansó and others (n35) 7.
[47] Gorwa (n12) 10.
[48] Sartor (n19) 43.
[49] ibid 51.
[50] See n5.
[51] Sartor (n19) 46.
[52] Frederik J Zuiderveen Borgesius, 'Strengthening legal protection against discrimination by algorithms and artificial intelligence' (2020) 24(10) The International Journal of Human Rights 1572, 1574.
[53] See in general Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2016).
[54] Gorwa (n12) 12.
[55] ibid 11. See also Celine Castets-Renard, 'Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement' (2020) U Ill JL Tech & Pol'y 283, 311.

extremely limited,[56] preventing removed lawful content from being reinstated.[57] Similar drawbacks are present in platforms' transparency reports,[58] which instead of providing the public with a critical understanding of how moderation works and therefore ensuring its accountability, look more like tools to ensure compliance with government censorship demands.[59] Furthermore, transparency can be hindered by copyright and trade secrets law.[60]

Ultimately, the lack of transparency prevents moderation decisions to be effectively contested and platforms themselves from being held accountable,[61] therefore impinging on due process, which is a right to a fair judicial process.[62]

1.2.3 Lack of due process

The absence of adequate transparency, contestability, and accountability, all cornerstones of due process,[63] prevents thus automated content moderation failures from being anticipated, pinpointed, and redressed,[64] leaving users without any 'realistic chance of challenging decisions made by platforms'.[65]

Having regard to the role platforms play in the public discourse, the growing use of automated filters without due process marks a significant shift in how speech regulation has traditionally been framed under the international human rights framework,[66] leading to a form of 'privatised speech regulation' that prioritises commercial interests over fundamental rights.[67] Ultimately, the lack of due process aggravates the impact of algorithmic moderation on citizens' fundamental rights – notably freedom of expression, freedom of information, and the right to privacy[68] – and transforms

[56] Maayan Perel and Niva Elkin-Koren, 'Accountability in algorithmic copyright enforcement' (2016) 19 Stanford Technology Law Review 473, 508.
[57] Mostert (n5) 381.
[58] For instance, the YouTube transparency report does not show how much of the removed content involved human review. See n41.
[59] Hannah Bloch-Wehba, 'Automation in Moderation' (2020) 53 Cornell Int'l LJ 41, 87-88.
[60] Perel (n56) 520-524.
[61] Frosio (n3) 28.
[62] ibid, Mostert (n5) 387.
[63] Frosio (n3) 28.
[64] Sartor (n19) 47.
[65] Nicolas Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4(3) Social Media and Society 1, 8 <https://journals.sagepub.com/doi/10.1177/2056305118787812> accessed 20 August 2021.
[66] Llansó (n33) 2.
[67] Cobbe (n15) 18.
[68] Frosio (n3) 13.

the digital arena into a space informed only by the pursuit of commercial interests of few global corporations.

In conclusion, while algorithms undoubtedly offer some advantages to platforms' moderating efforts, I argue that those are far outweighed by their chilling effect on fundamental rights, mainly driven by their lack of accuracy and due process, which can lead to what has been defined as 'digital prior restraint'[69] and a new form of 'corporate societal authority'.[70]

On the other hand, the amount of content to moderate and the necessity to protect users from the viral spread of illegal content[71] makes a 'human-only' moderation nearly impossible. For these reasons, I submit that while AI tools should play a role in online moderation, their use must be compliant with due process – in terms of transparency, contestability, and accountability – and thus respectful of users' fundamental rights, notably freedom of expression.

In the next chapter, I will discuss the current legal framework for content moderation, including proposals from international organisations and scholars to safeguard due process.

[69] Balkin (n8) 2017.
[70] Cobbe (n15) 5.
[71] Platforms' failure to address serious illegal content can lead to 'offline' violence and discrimination, as shown by the unaddressed incitement of violence that took place on Facebook against the Rohingya Muslim community in Myanmar. See UNGA, 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (9 October 2019) UN Doc A/74/486 para 41.

2. Legal framework governing algorithmic content moderation

In the previous chapter, I discussed several flaws of algorithmic content moderation, notably its inaccuracy, bias, and vulnerability. Yet, it was highlighted how its lack of transparency is the aspect of biggest concern, as it prevents moderation decisions from being effectively contested by users and platforms to be held accountable, therefore encroaching on due process and users' fundamental rights.

However, despite international organisations and scholars having long provided recommendations and guidelines to develop automated content moderation frameworks compliant with fundamental rights and due process, EU and national governments focussed instead on requiring or encouraging platforms to proactively moderate illegal content. Meanwhile, due process has been confined to self-regulation or soft law, incapable of changing the status quo described above, that is the complete absence of due process in content moderation.[72]

2.1 International framework

In a report issued in 2018, the UN Special Rapporteur on freedom of opinion and expression clearly stated that any regulation on AI should give pre-eminence to human rights, notably freedom of expression.[73] The report contains key recommendations for private companies to make their systems compliant by design to human rights.[74] First, 'radical' transparency must be upheld throughout the whole value chain to disclose to users the existence and functioning of AI processes (for content moderation, that includes information on removals and the outcome of redress procedures).[75] These measures would then enable companies to make effective remedy processes available to users, which should always include human review.[76] Finally, AI systems should undergo prior human rights impact assessments and external independent audits to ensure those transparency standards are met.[77]

[72] Perel (n56) 508, regarding copyright enforcement.
[73] UNGA, 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (29 August 2018) UN Doc A/73/348 para 43.
[74] ibid para 48.
[75] ibid paras 49-51.
[76] ibid para 60.
[77] ibid paras 53-55.

Similarly, the Committee of Ministers of the Council of Europe issued recommendations on the human rights impacts of algorithmic systems,[78] requiring private actors to respect human rights and embrace transparency, accountability, and contestability.[79] Furthermore, the document highlights the importance of allowing independent researchers to access datasets, to assess the impact of AI-driven services on users' rights and democracy.[80]

In the first chapter, I argued how the lack of transparency, as the backbone of due process, is the main drawback of automated moderation. In the author's opinion, then, the focus on transparency – and hence on contestability and accountability – in both the UN and Council of Europe documents is to be welcomed and shows how fragile fundamental rights are today in the context of automated moderation. In addition, I argue that independent auditing and external access to algorithms are key elements to ensure public oversight of platforms and therefore due process. Thus, they both should play a crucial role in any relevant regulation, including the DSA.

2.2 Scholars' proposals

Legal scholars too have developed rules to provide national legislators with guidance on how to design legal frameworks for online content moderation compliant with due process.

The well-established Manila[81] and Santa Clara principles,[82] for instance, provide useful provisions to enhance platforms' transparency, accountability, contestability, and public oversight.

Furthermore, published in 2021 and based on several principles suggested by Mostert,[83] the Aequitas principles offer a comprehensive set of rules on content moderation, which, unlike the Manila and Santa Clara principles, are meant to tackle both the over-blocking of user-generated content and the under-blocking of criminal content.[84] Moreover, they specifically address the issue of algorithmic moderation, highlighting the importance of allowing users to request human review

[78] Council of Europe, 'Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems' (8 April 2020) <https://rm.coe.int/09000016809e1154> accessed 20 August 2021.
[79] ibid 13.
[80] ibid 15.
[81] 'Manila Principles on Intermediary Liability' (2015) <https://manilaprinciples.org/> accessed 20 August 2021.
[82] 'The Santa Clara Principles on Transparency and Accountability in Content Moderation' (2018) <https://santaclaraprinciples.org/> accessed 20 August 2021.
[83] Mostert (n5) 388.
[84] 'Aequitas Principles on Online Due Process' (2021) <https://aequitas.online/principles/> accessed 20 August 2021. See **Appendix**.

over automated decisions[85] as well as setting tailored accountability, transparency, and non-discrimination obligations, including 'full disclosure' of platforms' moderation rules.[86]

All these scholarly initiatives manage to translate the fundamental rights and due process considerations discussed earlier into specific rules and should therefore form the basis of any regulation on content moderation. However, I submit that the Aequitas principles offer the most adequate framework on content moderation, capable of providing a flexible approach to solve its dilemma, that is finding a middle ground between under-blocking of illegal content and over-blocking of lawful content. Moreover, they provide tailored rules to safeguard due process and fundamental rights in the context of algorithmic moderation.

On the other hand, I argue that the 'expedited removal' of the most serious criminal content suggested by the Aequitas principles,[87] although understandable in principle, might be problematic. Having regard to the difficulty of developing definitions capable of effectively recognising the difference between lawful or unlawful content,[88] I submit that entrusting private platforms with the task of categorising illegal content can only lead to over-blocking or arbitrariness. Therefore, I submit that 'fast-track' procedures should only be allowed pursuant to court orders.

2.3 EU and national initiatives

While international and scholars' recommendations are both guided by the respect of human rights, recent EU and national initiatives do not seem to go in the same direction.

In the first chapter, I highlighted how the prohibition of general monitoring obligations laid down in the ECD[89] has been put under pressure by subsequent sectoral legislation, that introduced proactive obligations much close to *ex-ante* filtering.[90] Apart from such instruments,[91] automated content moderation has been regulated in the EU mainly through soft law or self-regulation aimed at

[85] ibid para 4.4.
[86] ibid para 4.6.
[87] ibid para 3.
[88] Sartor (n19) 57.
[89] See n27.
[90] See n32. See also Giancarlo Frosio, 'The Death of 'No Monitoring Obligations': A Story of Untameable Monsters' (2017) 8 JIPITEC 199 <https://www.jipitec.eu/issues/jipitec-8-3-2017/4621/JIPITEC_8_3_2017_199_Frosio> accessed 20 August 2021.
[91] The EU legislative framework also includes rules applicable to certain illegal content, addressed to Member States rather than platforms, built on the ECD notice-and-takedown architecture. See De Streel (n2) 15-32.

encouraging online intermediaries to adopt voluntary proactive measures against illegal content, with scarce emphasis on due process.

For instance, the Code of Conduct against the spread of hate speech online[92] contains several commitments aimed, among others, at reviewing notices on such content – and eventually remove it – in less than 24 hours.[93] However, while the evaluations highlight the rate and speed of removals,[94] little is known about whether the removed content was indeed illegal or the outcome of any redress mechanisms.[95] Moreover, the Code fails to address due process concerns related to the use of automated measures, which are not mentioned at all.[96] These weaknesses have led commentators to view the Code as an instrument mainly incentivising private censorship and over-blocking.[97]

Algorithmic enforcement was also specifically mentioned in the Communication of 2017 on tackling illegal content online,[98] where the Commission strongly favoured the use of automated proactive measures to detect illegal content, clarifying how such behaviour does not automatically lead to the loss of the liability exemption laid down in Art. 14 ECD.[99] Although the Communication mentions the importance to avoid over-blocking and respect fundamental rights, the lack of sanctions for non-compliance, together with the arbitrariness of platforms' ToS, might lead intermediaries to preventively take down content to avoid liability.[100]

Furthermore, automated proactive measures are encouraged in the Recommendation issued in 2018 as a follow-up to the previous Communication,[101] although here the Commission was more

[92] European Commission, 'Code of Conduct on countering illegal hate speech online' (*European Commission*, 30 June 2016) <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 20 August 2021.
[93] De Streel (n2) 30.
[94] Directorate-General for Justice and Consumers, 'Countering illegal hate speech online 5th evaluation of the Code of Conduct' (*European Commission*, June 2020) <https://ec.europa.eu/info/sites/default/files/codeofconduct_2020_factsheet_12.pdf> accessed 20 August 2021.
[95] Teresa Quintel and Carsten Ullrich, 'Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond', in Bilyana Petkova and Tuomas Ojanen (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar 2019) 206.
[96] ibid 205.
[97] De Streel (n2) 31.
[98] European Commission, 'Tackling Illegal Content Online. Towards an enhanced responsibility of online platforms' (Communication) COM(2017) 555 final.
[99] ibid para 3.3.2.
[100] Quintel (n95) 208-209.
[101] Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online [2018] OJ L63/50.

ambitious in terms of transparency and accountability,[102] highlighting the need for automated measures to be accompanied by 'effective and appropriate safeguards'.[103] However, while the Recommendation contains several such safeguards to avoid the unjust removal of content, such provisions are too vague to be effective.[104] Moreover, I argue that since the Recommendation is without prejudice to the enforcement of platforms' ToS,[105] the effectiveness of those safeguards is further watered down.

It is not surprising, then, that commentators argued how such a political strategy puts the protection of fundamental rights to a difficult test.[106] In the author's opinion, instead of balancing moderation's lack of due process with relevant obligations for platforms, the EU preferred to prioritise the swift removal of (alleged) illegal content mainly for the benefit of rightsholders and platforms' traditional competitors. In doing so, the safeguard of due process and fundamental rights has been left to voluntary self-regulation, with no sanctions for non-compliance.

Fundamental rights in the context of algorithmic moderation find little attention also at the national level.

Legislation adopted in France[107] and Germany[108] was mentioned in the first chapter. While the former was declared unconstitutional for its incompatibility with freedom of expression,[109] the latter, although not being the censorship tool it was early depicted as,[110] seems to have failed at introducing due process, leaving users without effective redress mechanisms.[111]

In the UK, the recent White Paper on Online Harms[112] also scarcely addressed freedom of expression and due process.[113] However, a report[114] issued after the consultation phase sought to give more

[102] Quintel (n95) 208-209.
[103] Commission Recommendation (n101), para 18.
[104] Quintel (n95) 210.
[105] Commission Recommendation (n101), para 19.
[106] Frosio (n3) 7.
[107] See n24.
[108] See n23.
[109] See n25.
[110] Thomas Wischmeyer, 'What is illegal offline is also illegal online: the German Network Enforcement Act 2017' in Bilyana Petkova and Tuomas Ojanen (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar 2019) 55.
[111] ibid 49.
[112] Secretary of State for Digital, Culture, Media & Sport and Secretary of State for the Home Department, *Online Harms White Paper* (CP 57, 2019).
[113] Mostert (n5) 385.
[114] Department for Digital, Culture, Media & Sport and Home Office, 'The government report on transparency reporting in relation to online harms' (*GOV.UK*, 15 December 2020)

attention to transparency, by stressing, for instance, that the quantity of content removed should not be the focus of reporting, given its potential perverse effect of encouraging over-blocking.[115] On the other hand, I argue that the calls for the safeguard of the commercially sensitive nature of algorithms[116] could hinder accountability and public oversight.

Indeed, building on the conceptual framework outlined in the White Paper, the UK Government recently published the draft Online Safety Bill,[117] which looks like the UK version of the DSA. However, unlikely the latter, the former would place general filtering obligations on all the platforms within its scope, thus representing a landmark departure from the ECD.[118] Moreover, against the risks of 'collateral damage' to freedom of expression arising from these obligations, the draft Bill is viewed as lacking effective due process safeguards,[119] making the proposed legal framework a 'recipe for censorship'.[120]

In the author's opinion, the strategy put forward by the EU and national instruments described above, that is to require intermediaries to proactively engage in content moderation without effective due process obligations, has ultimately entrusted online speech regulation and freedom of expression to private actors.

Therefore, self-regulation should be abandoned in favour of overarching horizontal legislation underpinned by mandatory due process obligations, as recommended by international organisations and scholars' initiatives, among which the Aequitas principles are the most persuasive. I argue that such a regulatory framework should apply not only to illegal content but also to that contrary to platforms' ToS, which are often stricter than national laws and cause over-blocking.[121]

<https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/government-transparency-report> accessed 20 August 2021.

[115] ibid Recommendation 23.

[116] ibid Recommendations 30, 35.

[117] Minister of State for Digital and Culture, *Draft Online Safety Bill* (CP 405, 2021).

[118] Graham Smith, 'Harm Version 3.0: the draft Online Safety Bill' (*Informm*, 1 June 2021) <https://inforrm.org/2021/06/01/harm-version-3-0-the-draft-online-safety-bill-graham-smith/#more-49278> accessed 20 August 2021, analysing s9(3) of the draft Bill.

[119] Christoph Schmon, 'UK's Draft Online Safety Bill Raises Serious Concerns Around Freedom of Expression' (*EFF*, 14 July 2021) <https://www.eff.org/it/deeplinks/2021/07/uks-draft-online-safety-bill-raises-serious-concerns-around-freedom-expression> accessed 20 August 2021.

[120] Alex Hern, 'Online safety bill 'a recipe for censorship', say campaigners' (*The Guardian*, 12 May 2021) <https://www.theguardian.com/media/2021/may/12/uk-to-require-social-media-to-protect-democratically-important-content> accessed 20 August 2021.

[121] De Streel (n2) 43.

Based on these premises, I submit that the DSA is a historical chance to build a digital space finally safeguarding due process and freedom of expression. The next chapter will analyse the Commission proposal and propose amendments to overcome some of its shortcomings and achieve its due process ambitions.

3. The Digital Services Act

Announced as one of the key actions of President von der Leyen to ensure an 'open, democratic and sustainable society',[122] the DSA could represent a landmark change in the EU regulatory framework of online intermediaries. Aiming at revising 20-year-old rules on platforms' liability and imposing new obligations on due process,[123] I submit that the DSA represents a welcomed political step back from the previous EU soft law approach. Furthermore, by choosing a Regulation instead of a Directive, the Commission finally recognises the seriousness of the impact of algorithmic moderation on users' fundamental rights, which can only be addressed by harmonised rules at the EU level.

This chapter discusses the most important DSA provisions concerning automated moderation, which are divided into three main thematic areas – scope, liability, and due process obligations[124] (broken down into transparency, contestability, and accountability measures[125]) – and proposes amendments thereto. Reference will be made, when relevant, to the state of play of the EU negotiations, notably by commenting on the European Parliament Rapporteur's draft report[126] and the Council Presidency's compromise text.[127]

3.1 Scope

This section focuses on the nature of the content within the scope of the DSA and the relation between the latter and EU sectoral legislation.

[122] European Commission, 'Shaping Europe's digital future' (Communication) COM(2020) 67 final 1.
[123] João P Quintais and Sebastian F Schwemer, 'The Interplay between the Digital Services Act and Sector Regulation: How Special is Copyright?' (*SSRN*, 10 May 2021) (draft) 1 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3841606> accessed 20 August 2021.
[124] The DSA defines them as 'due diligence obligations'.
[125] Some provisions might indeed overlap between these three elements.
[126] European Parliament, IMCO Committee, 'Draft report 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' PE693.594v01-00 (28 May 2021) <https://www.europarl.europa.eu/doceo/document/IMCO-PR-693594_EN.pdf> accessed 20 August 2021.
[127] Council of the European Union, '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 - Presidency compromise text on Chapters II, IV and V, with respective recitals' 9288/1/21 REV (16 June 2021) <https://data.consilium.europa.eu/doc/document/ST-9288-2021-REV-1/en/pdf> accessed 20 August 2021.

3.1.1 Nature of the content

Like the ECD, the DSA only addresses illegal content, defined as 'any information […] not in compliance with Union law or the law of a Member State',[128] thus leaving out of its scope the notion of harmful content. Whether or not regulating the latter has been extensively debated over the years: while several commentators argued that obligations on harmful content would be detrimental to users' freedom of expression,[129] others submitted that the lack of rules on such category further increases platforms' discretion and censorship powers.[130]

I argue that the Commission correctly excluded harmful content from the scope of the DSA. Being a vague and open-ended category, which can considerably vary across Member States, the inclusion of harmful content would empower platforms with too much discretion, especially with the use of automated means, as well as seriously undermine users' right to redress. Besides, self-regulatory measures on disinformation, a prime example of harmful content, have been shown to increase over-blocking and private censorship.[131] Thus, I submit that the absence of references to harmful content in both the IMCO's and Council's documents is to be welcomed.

In the author's opinion, the DSA also correctly does not provide for different scenarios based on the seriousness of illegal content. On the other hand, the IMCO Draft Report proposes a 'fast track' procedure – removals within 24 hours – for that illegal content apt to 'seriously harm public policy, public security or public health or seriously harm consumers' health or safety'.[132]

As discussed in the previous chapter, I argue that qualitative assessments of illegal content should only be left to judicial bodies, instead of private companies. Moreover, short deadlines lead platforms to sacrifice accuracy to show compliance, not only in the context of removals but also in that of their transparency reports.[133]

The same reasoning applies a fortiori in the context of the DSA – which covers the whole spectrum of EU and national illegal content – where references to the seriousness of illegal content would

[128] DSA, art 2(g). Note the improvement from the ECD, which failed to define 'illegal activities'.
[129] De Streel (n2) 78; Andrea Bertolini, Francesca Episcopo and Nicoleta-Angela Cherciu, *Liability of online platforms* (European Parliament 2021) 80.
[130] Ethan Shattock, 'Self-regulation 2:0? A critical reflection of the European fight against disinformation' (2021) 2(3) Harvard Kennedy School Misinformation Review 4 <https://misinforeview.hks.harvard.edu/wp-content/uploads/2021/05/shattock_self_regulation_european_disinformation_20210531.pdf> accessed 20 August 2021.
[131] Quintel (n95) 213.
[132] IMCO (n126) am 71.
[133] See n95.

require assessments on a tremendously diversified range of content, resulting at best in discretionary choices by platforms, and at worst in impossible ones, especially for SMEs.

Therefore, I submit that 'fast-track' procedures should continue to be excluded from the scope of the DSA or, if eventually included, only allowed in compliance with a court order.

3.1.2 Interplay with sectoral legislation

In its Article 1(5), the DSA states that the Regulation is without prejudice to sectoral EU legislation, like the AVMSD and the Copyright Directive, which are *lex specialis* to the DSA. In the Commission's view, the DSA should indeed only apply to circumstances not covered by such acts or in cases where Member States kept a degree of flexibility.[134]

However, Quintais and Schwemer suggest how the boundaries between the DSA and Article 17 Copyright Directive are not as obvious as advocated by the Commission, with several 'grey areas' open to interpretation.[135] In the author's opinion, similar considerations should apply to Article 28b AVMSD.[136]

Since the DSA should become the future baseline regime of EU law, I submit that as suggested by Quintais and Schwemer, the Regulation should prevail over sectoral legislation, unless the latter expressly provides for more precise rules.[137]

3.2 Intermediaries' liability

Discussions on changing ECD's rules on online intermediaries' liability have been multiplied over the last decade, due to the massive escalation of platforms' size, turnover, and influence. With the DSA proposal, the Commission chose to maintain the legal architecture of the ECD – whose relevant content is restated in Articles 3, 4, 5, and 7 DSA – in light of the principles that emerged from the EUCJ case law.[138]

For the purposes of this paper, it suffices to note that Article 5 DSA replicates the ECD's knowledge-based liability exemption and the notice-and-take-down mechanism for hosting service providers –

[134] DSA, Recital 9.
[135] Quintais (n123) 21.
[136] See n30.
[137] Quintais (n123) 21.
[138] Caroline Cauffman and Catalina Goanta, 'A New Order: The Digital Services Act and Consumer Protection' (2021) 00 European Journal of Risk Regulation 1, 6.

which include online platforms[139] – so that those are not liable for the content they store, if they do not have actual knowledge of its illegal nature or when they know, they act expeditiously to remove that content. Importantly, Article 7 DSA restates the prohibition of general monitoring obligations set out in Article 15 ECD.

3.2.1 Knowledge-based liability and prohibition of general monitoring obligations

Commentators have been split over whether maintaining or changing the ECD's liability exemption and general monitoring ban. Favourable to the former option, viewed as the best suited to safeguard users' fundamental rights, are scholars like Frosio[140] and several civil society organisations.[141] Similarly, Bayer argues that the notice-and-take-down procedure should apply only to manifestly illegal content, leaving the less intrusive notice-and-notice procedure as the default regime.[142]

On the other hand, a change of ECD's rules on liability is called on by quite a few commentators. Sartor and Loreggia, for instance, suggest the introduction of proactive obligations on platforms, which should adopt 'reasonable measures' to prevent or mitigate grave harm, the failure of which, if resulting in actual harm, should then trigger liability.[143] Smith goes even further, arguing for the abandonment of both the liability exemption and the general monitoring ban, which he considers as prioritising free speech – at least platforms' interpretation thereof – at the detriment of other EU fundamental rights.[144]

In the author's opinion, both the ECD's liability exemption – built on a notice-and-take-down procedure – and the prohibition of general monitoring obligation must be maintained. As discussed earlier, the increasing use of automated tools in content moderation has already put private companies in charge of speech regulation worldwide. Considering the vital role platforms play as

[139] DSA, art 2(h).
[140] Frosio (n3) 4.
[141] Chloé Berthélémy and Jan Penfrat, 'Platform Regulation Done Right - EDRi Position Paper on the EU Digital Services Act' (*EDRi*, 9 April 2020) 17-18 <https://edri.org/wp-content/uploads/2020/04/DSA_EDRiPositionPaper.pdf> accessed 20 August 2021. See also Article 19, 'At a glance: Does the EU Digital Services Act protect freedom of expression?' (*Article 19*, 11 February 2021) <https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/> accessed 20 August 2021.
[142] Under such a regime, the provider who receives a notice about a content deemed illegal should forward it to the user concerned, instead of being required to remove that content. See Judit Bayer, *Between Anarchy and Censorship – Public discourse and the duties of social media* (CEPS 2019) 15, 20.
[143] Sartor (n19) 61-62.
[144] Melanie Smith, *Enforcement and cooperation between Member States – E-Commerce and the future Digital Services Act* (European Parliament 2020) 31-33.

public spaces, I argue that any tightening of the ECD rules to increase providers' liability would inevitably lead to more over-blocking of lawful content and private algorithmic censorship, with inadmissible interferences to fundamental rights.

3.2.2 When should liability arise?

Other than from its own-initiative investigations[145] or judiciary and administrative authorities,[146] hosting providers' knowledge and liability can arise from users' notices reporting alleged illegal content, when compliant with Article 14(2) DSA.[147]

Several civil society organisations advocate that intermediaries' knowledge should be triggered only after the receipt of a court order,[148] while users' notices should only prompt notice-and-notice 'plus' mechanisms, which would oblige hosting providers to remove the alleged illegal content only in case its author does not submit any counter-notice.[149]

In the author's opinion, while tying hosting providers' knowledge only to the receipt of court orders is theoretically persuasive and would reduce over-blocking, the virality and volume of illegal content online would make such a form of moderation practically unfeasible, with vast portions of illegal content bound to go unnoticed. At the same time, it is recognised that the current wording of Article 14(3) DSA is still problematic for freedom of expression. Thus, some alternatives are suggested.

First, instead of all notices, knowledge could be triggered – other than by judicial orders – by those orders sent only by trusted flaggers as defined in Article 19 DSA, which can guarantee a heightened level of competence and accuracy compared to average users. Alternatively, the notice-and-take-down framework could apply only to serious illegal content, to be defined in the Regulation.

In both cases, users' notices should only prompt a notice-and-notice 'plus' mechanism.[150]

---

[145] DSA, Recital 22.
[146] DSA, art 8.
[147] DSA, art 14(3). The system of notices will be further discussed below.
[148] EFF, 'Preserve What Works, Fix What is Broken: EFF's Policy Principles for the Digital Services Act' (*EFF*, 2020) 6 <https://www.eff.org/files/consolidatedeupolicyprinciples.pdf> accessed 20 August 2021.
[149] Access Now, 'Access Now's Position on the Digital Services Act Package' (*Access Now*, September 2020) 4-5 <https://www.accessnow.org/cms/assets/uploads/2020/10/Access-Nows-Position-on-the-Digital-Services-Act-Package.pdf> accessed 20 August 2021. The difference with a normal 'notice-and-notice' procedure is that the latter only requires providers to forward the notice to the author of the alleged illegal content, without being obliged to remove it. See n142.
[150] ibid

3.2.3 A 'quasi-Good Samaritan' clause

The main novelty of the DSA in terms of liability is offered by its Article 6, whereby providers do not lose the liability exemption solely because they carry out voluntary investigations aimed at detecting illegal content. This provision, which codifies the statement of the Commission in its Communication from 2017,[151] clearly seeks to introduce into the EU law a 'Good Samaritan' protection for online intermediaries.[152]

However, while the original US version exempts intermediaries from liability when they proactively remove illegal content and also when they fail to do so,[153] Article 6 DSA covers only the former part, leaving providers liable in case they do not expeditiously remove illegal content after having obtained knowledge thereof.[154] Kuczerawy points out the contradictory nature of such wording, which could either lead platforms to restrain from any proactive moderation of illegal content or, on the other hand, over-block in order not to lose liability.[155]

In the author's view, proactive monitoring of content should not be mandated nor encouraged by the legislator,[156] since that would likely result in massive use of algorithms, with their well-known adverse effects on users' fundamental rights. Therefore, I suggest deleting Article 6 DSA altogether, in line with the recommendation of the Opinion of the European Parliament LIBE Committee.[157]

3.3 Due process obligations

The DSA proposal contains several provisions aimed at ensuring due process in content moderation. Interestingly, the Commission avoided a 'one-size-fits-all' approach but rather established a set of asymmetric obligations based on the type and size of intermediaries. After some general rules

[151] See n99.
[152] Communications Decency Act of 1996, 47 U.S.C. § 230.
[153] Joan Barata, 'Positive Intent Protections: Incorporating a Good Samaritan principle in the EU Digital Services Act' (*Center for Democracy & Technology*, 29 July 2020) <https://cdt.org/insights/positive-intent-protections-incorporating-a-good-samaritan-principle-in-the-eu-digital-services-act/> accessed 20 August 2021.
[154] Aleksandra Kuczerawy, 'The Good Samaritan that wasn't: voluntary monitoring under the (draft) Digital Services Act' (*Verfassungsblog*, 12 January 2021) <https://verfassungsblog.de/good-samaritan-dsa/> accessed 20 August 2021.
[155] ibid
[156] For these reasons, I argue that the European Parliament Rapporteur's addition of vague 'appropriate safeguards' is still insufficient. See IMCO (n126) am 75.
[157] European Parliament, LIBE Committee, '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' PE692.898v06-00 (28 July 2021) am 62 <https://www.europarl.europa.eu/doceo/document/LIBE-AD-692898_EN.pdf> accessed 20 August 2021.

applicable to all providers, the DSA introduces specific rules addressed to hosting providers and online platforms,[158] with additional obligations for very large online platforms ('VLOPs').[159]

This section breaks up DSA's due process obligations into transparency, contestability, and accountability measures.

3.3.1 Transparency

The DSA sets out different provisions to enhance intermediaries' transparency.

First, Article 12 DSA requires all intermediaries to provide users with information on content moderation decisions made according to their ToS, including any use of algorithmic tools. I argue that this requirement is crucial since ToS are often stricter than national laws.[160] Article 12(2) DSA also requires such voluntary content moderation to be compliant with the fundamental rights set out in the EU Charter of Fundamental Rights.

The DSA introduces then an obligation for providers to produce annual reports on any content moderation they engage in, whether voluntary or mandated by the DSA itself. While Article 13 contains minimum reporting obligations applicable to all providers,[161] Articles 23 and 33 list additional elements to be disclosed by online platforms and VLOPs respectively.

First, Article 23 DSA requires online platforms to disclose information on out-of-court dispute settlements, suspensions following misuse, and any use of automated moderation tools. Second, VLOPs must also include in their biannual reports information on risk assessments and audits.[162]

I submit that the DSA's focus on disaggregated data – necessary for meaningful reports[163] – represents a clear step up from previous approaches,[164] whose scarce reporting obligations appear to be more concerned with the outcome of moderation rather than its reliability.[165]

[158] For the definition, see n1. As per art 16 DSA, rules addressed to online platforms do not apply to micro and small enterprises.
[159] Very large platforms are defined by art 25(1) DSA as having at least 45 million average monthly users in the EU.
[160] See n121.
[161] All reports must include quantitative and qualitative information on orders and notices received as well as on the outcome of internal complaint mechanisms.
[162] DSA, art 33.
[163] Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation' (2020) 43 Fordham Int'l LJ 939, 992.
[164] For instance, neither the AVMSD nor the Copyright Directive contains similar provisions.
[165] See n95.

On the other hand, transparency around the use of algorithms seems still inadequate. To begin with, I argue that all intermediaries – not only hosting providers – should include in their reports information on the use of automated means. That requirement should thus be moved from Article 23(c) to Article 13 DSA.

Furthermore, information on the use of algorithms should be more fine-grained. It is paramount that all intermediaries provide not only information on the number of content flagged and removed through algorithmic means, but also on whether and at what stage human review took place.

3.3.2 Contestability

Starting from the ECD regime, the DSA seeks to further specify and harmonise the notice-and-action procedure at the EU level. By setting out procedural rules on notices, statements of reasons, and redress mechanisms, the Commission finally recognises that any content moderation framework cannot disregard due process.

3.3.2.1 Notice-and-action procedure

Article 14 DSA requires hosting providers to allow users to submit notices reporting alleged illegal content that, if including certain minimum information, trigger providers' liability. Article 15 DSA introduces then a statement of reasons that providers must send to users whose content has been removed – because it was either illegal or contrary to ToS – to inform them of the reasons for such removal[166] and of any available redress mechanisms. Finally, Article 19 DSA introduces trusted flaggers, while Article 20 DSA provides for measures against misuse.

To begin with, I argue that given the social role played by platforms, any removal of content allegedly contrary to ToS should not take place before the concerned user had a chance to submit a counter-notice.[167]

Next, Article 15 DSA is a key provision, since informative statements accompanying removals are essential to guarantee due process, notably the transparency and contestability of platforms' decisions. However, Article 15(2)(c) DSA, which aims at providing users with information on the use

[166] Information includes how and why the content was removed and, notably, whether the decision was made through automated means.
[167] See n150.

of automated means, should be reinforced, with clear information on the logic and reasons behind such decisions, considering the binary nature – 'ignore' or 'delete' – of algorithms.[168]

Then, I submit that the introduction of 'trusted flaggers' by Article 19 DSA, with clear requirements for the award and loss of their status by Digital Service Coordinators,[169] enhances the accuracy, accountability, and transparency of moderation,[170] and it should thus be supported.

Finally, I argue that the temporary suspension of accounts and processing of notices due to the frequent submission of manifestly illegal content or unfounded notices, respectively, is a proportionate measure to discourage abuses.[171] However, I submit that the suspension of the processing of users' complaints should never be allowed,[172] as it would deprive users of their fundamental right to a due process, especially given the imbalance of power between them and platforms.

3.3.2.2 Redress mechanisms

Articles 17 and 18 DSA introduce the most important elements for a content moderation framework respectful of due process, that is the right for users to contest platforms' decisions and start an appeal procedure.

Article 17 DSA allows users to lodge internal complaints against moderation decisions, correctly specifying that if the complaint is successful, platforms must reinstate the removed content or reactivate the users' account.[173]

In the author's opinion, such provision succeeds overall in increasing due process in content moderation. However, I argue that Article 17(5) DSA, whereby platforms' appeal decisions should not 'solely' be taken by automated means, is not enough. On the contrary, I submit that the DSA should be explicit in ensuring users have always the right to subject their complaints to human review. Indeed, given the inaccuracy of algorithms, only a human moderator can offer an

[168] Giovanni De Gregorio, 'Democratising online content moderation: A constitutional framework' (2020) 36 Computer Law & Security Review 1, 15.
[169] DSA, art 39.
[170] Castets-Renard (n55) 319.
[171] DSA, art 20.
[172] DSA, art 20(2).
[173] DSA, art 17(3).

appropriate level of due process in reviews. Furthermore, clear deadlines should be imposed on platforms' internal reviews, as proposed in the IMCO Draft Report.[174]

Finally, Article 18 DSA allows users to activate out-of-court dispute mechanisms – thus going beyond the mere possibility provided for by Article 17 ECD – to settle any dispute not resolved by internal appeals. I argue that the requirements of independence, expertise, and fairness of the dispute settlement bodies, to be certified by Member States, are appropriate to overcome the pitfalls of 'made-up corporate PR tool[s]'[175] such as Facebook's Oversight Board.[176]

3.3.3 Accountability

This section focuses on three key provisions to enhance the accountability of automated moderation, all addressed to VLOPs, that is risk assessments, audits, and data access.

Article 26 DSA requires VLOPs to conduct yearly self-assessments on systemic risks that might arise from the use of their services. Such provision, welcomed in principle, is in line with calls from international organisations,[177] notably where it focuses the assessments on 'negative effects for the exercise of the fundamental rights'.[178] However, the requirement for VLOPs to adopt effective mitigation measures to address the systemic risk of the dissemination of illegal content could result in the circumvention of DSA's transparency and due process measures.[179] Therefore, I submit that Article 26(1)(a) should be deleted.

Article 28 DSA requires then VLOPs to be subject to independent audits to assess their compliance with the due process obligations analysed above. Since audits are key to ensure the transparency and accountability of content moderation, such provision is critical and should thus be supported. However, I submit that the respect of fundamental rights should be an express object of audits.

Finally, Article 31 DSA introduces another pivotal provision to ensure VLOPs' algorithmic accountability, that is the access to their data by the relevant Digital Services Coordinator, the Commission, and vetted researchers.

---

[174] IMCO (n126) am 101.
[175] As per Jesse Lehrich, quoted in Hannah Murphy, 'Facebook's Oversight Board: an imperfect solution to a complex problem' (*Financial Times*, 17 May 2021) <https://www.ft.com/content/802ae18c-af43-437b-ae70-12a87c838571> accessed 20 August 2021.
[176] Bloch-Wehba (n59) 92-93.
[177] See n77.
[178] DSA, art 26(1)(b).
[179] DSA, arts 26(1)(a), 27(1)(a) and Recital 58.

I argue that data access by researchers is one of the most important features of the DSA, as it allows public and independent scrutiny not only of moderation algorithms but also indirectly of VLOPs' risk assessments and audits.[180] On the other hand, its effectiveness could be hindered by the strict requirements on researchers and by trade secrets.

First, the requirement for researchers to be able to preserve the security of data might prove difficult to achieve, especially for large datasets. One possible solution could be for the Commission to establish a centralised portal to allow researchers to securely access data.[181]

Second, trade secrets could block or limit data access.[182] Certainly, such a legal barrier would be overcome if platforms were obliged to use open-source algorithms for content moderation, as emphasised by Mostert and Urbelis.[183] In the author's opinion, however, while this proposal would certainly tackle several of the transparency and accountability issues described in this paper, it would be politically almost impossible to achieve. Thus, the DSA should at least stress that only duly substantiated requests – to be proved by VLOPs – can allow them to refuse access to data based on confidential information. Besides, that of trade secrets might even be an ill-founded problem. Indeed, technical solutions already exist to check data while keeping the source code undisclosed.[184] Moreover, the purpose of data access is to disclose the result of automated moderation, rather than algorithms themselves.[185]

Finally, data access could be hampered by another legal barrier, which is the GDPR. When providing researchers access to data, platforms would indeed be subject, as controllers, to several GDPR obligations.[186] While the sharing of data to researchers might be compliant with Article 89 (1) GDPR, I agree with Vermeulen that the Commission and stakeholders should develop codes of conduct to specify the interplay between the GDPR and Article 31 DSA.[187]

---

[180] Alex Engler, 'Platform data access is a lynchpin of the EU's Digital Services Act' (*Brookings*, 15 January 2021) <https://www.brookings.edu/blog/techtank/2021/01/15/platform-data-access-is-a-lynchpin-of-the-eus-digital-services-act/> accessed 20 August 2021.
[181] ibid
[182] DSA, arts 31(4) and 31(6)(b).
[183] Frederick Mostert and Alex Urbelis, 'Social media platforms must abandon algorithmic secrecy' (*Financial Times*, 17 June 2021) <https://www.ft.com/content/39d69f80-5266-4e22-965f-efbc19d2e776> accessed 20 August 2021.
[184] Joshua Kroll, 'Accountable Algorithms (A Provocation)' (*Media@LSE*, 10 February 2016) <https://blogs.lse.ac.uk/medialse/2016/02/10/accountable-algorithms-a-provocation/> accessed 20 August 2021.
[185] Castets-Renard (n55) 319.
[186] Mathias Vermeulen, 'The Keys to the Kingdom. Overcoming GDPR concerns to unlock access to platform data for independent researchers' (*Knight First Amendment Institute*, 27 July 2021) <https://knightcolumbia.org/content/the-keys-to-the-kingdom> accessed 20 August 2021.
[187] ibid

Summary of Findings and Conclusion

Given their role played in society, online intermediaries should engage in the moderation of content to ensure their users are protected from illegal activities while being free to share lawful content. However, this paper demonstrated how content moderation, now mainly driven by algorithms, not only is often inaccurate and discriminatory but also lacks due process and thus undermines users' fundamental rights. Notably, its absence of transparency prevents moderation decisions from being contested and, ultimately, platforms from being held accountable.

Next, I argued how this scenario has been further exacerbated by a legal framework that prioritised the swift removal of illegal content through the imposition of proactive obligations, leaving the safeguard of due process to (mostly ineffective) self-regulation and soft law. I thus submitted that the recent DSA proposal is a promising attempt to establish an overarching framework on content moderation inspired by due process principles, as long advocated by international organisations and scholars.

I then analysed the DSA proposal and suggested amendments to further safeguard due process and users' fundamental rights, especially restricting the circumstances when platforms' liability can arise and strengthening algorithmic transparency and accountability, which still lack in the proposal.

In conclusion, I submit that the DSA is a historical chance to shape a 'due process by design' digital space whose rules are set by democratic institutions and not private actors, and where users can fully enjoy their fundamental rights of freedom of expression and speech. Moreover, like the GDPR,[188] the DSA can go well beyond the EU borders, paving the way for a global regulation of online content moderation underpinned by human rights principles,[189] notably when the United States is also considering introducing stricter rules for platforms.[190]

Recommendations

Based on the analysis carried out in Chapter 3, the following amendments and considerations are recommended to the European Commission's DG Connect, with a view to the upcoming trilogues

[188] Ansgar Koene and others, *A governance framework for algorithmic accountability and transparency* (European Parliamentary Research Service 2019) 75.
[189] Damian Tambini, 'Media Policy in 2021. As the EU takes on the tech giants, will the UK?' (*Media@LSE*, 12 January 2021) <https://blogs.lse.ac.uk/medialse/2021/01/12/media-policy-in-2021-as-the-eu-takes-on-the-tech-giants-will-the-uk/> accessed 20 August 2021.
[190] Kiran Stacey and Hannah Murphy, 'Now Republicans and Democrats alike want to rein in Big Tech' (*Financial Times*, 12 January 2021) <https://www.ft.com/content/e7c1a64f-b2d9-423b-a86c-f36d1c4e71b7> accessed 20 August 2021.

with the European Parliament and the Council, to make the DSA further compliant with due process and fundamental rights.

Scope

i. Harmful content should continue to be excluded from the scope of the DSA.
ii. If ever included in the DSA, 'fast-track' procedures should only be allowed in compliance with a court order.
iii. The DSA's horizontal obligations should apply to intermediaries covered by sectoral legislation unless the latter expressly provides for stricter rules.

Liability

i. Intermediaries' knowledge-based liability exemption (Articles 3-5 DSA) and the prohibition of general monitoring obligations (Article 7 DSA) should be retained.
ii. Amendment to Article 14(3) DSA: hosting providers' knowledge should be triggered only by judicial authorities' orders and trusted flaggers' notices or only for serious, defined illegal content. Notice-and-notice 'plus' mechanisms should apply to users' notices.
iii. Article 6 DSA should be deleted, as it could further increase over-blocking.

Transparency

i. Article 15(2)(c) DSA should require hosting providers to inform users about the logic and reasons behind automated moderation decisions.
ii. Information on the use of automated means of moderation should be moved from Article 23(c) to Article 13 DSA, applicable to all intermediary providers.
iii. Information on the use of algorithms required by Article 23(c) DSA should include information on whether and at what stage human review took place.

Contestability

i. Any removal of content contrary to intermediaries' ToS should not take place before the concerned user had a chance to submit a counter-notice.
ii. Amendment to Article 17(5) DSA: users should always have the right to subject their complaints to human review. Furthermore, clear deadlines should be imposed on platforms' internal reviews.

iii. Amendment to Article 20(2) DSA: the suspension of the processing of users' complaints should never be allowed, as it would be detrimental to due process.

Accountability

i. Article 26(1)(a) DSA should be deleted, to avoid any risks of circumvention of DSA's due process obligations.
ii. Amendment to Article 28 DSA: the compliance of VLOPs' moderation with fundamental rights should be an express object of audits.
iii. Amendments to Article 31 DSA:
    a. A centralised portal should be developed to allow researchers to securely access data.
    b. Only duly substantiated requests, to be proved by VLOPs, should allow them to refuse access to data based on confidential information.
    c. The interplay between the DSA and the GDPR in the context of data access should be clarified, for instance through the development of codes of conduct.

Bibliography

Table of legislation

European Union

Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online [2018] OJ L63/50

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

Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services [2018] OJ L 303

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

France

Loi n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet

Germany

Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act)

United States

Communications Decency Act of 1996, 47 U.S.C. § 230

Other Sources

A

Access Now, 'Access Now's Position on the Digital Services Act Package' (*Access Now*, September 2020) <https://www.accessnow.org/cms/assets/uploads/2020/10/Access-Nows-Position-on-the-Digital-Services-Act-Package.pdf> accessed 20 August 2021

Access Now and others, 'Civil Society Statement of Key Principles – Committee on Civil Liberties, Justice and Home Affairs (LIBE) Report on the draft Digital Services Act' (*Center for Democracy & Technology*, 8 June 2021) <https://cdt.org/wp-content/uploads/2021/06/2021-06-08-CDT-Europe-Joint-Statement-in-Advance-of-the-LIBE-Vote.pdf> accessed 20 August 2021

'Aequitas Principles on Online Due Process' (2021) <https://aequitas.online/principles/> accessed 20 August 2021

Article 19, 'At a glance: Does the EU Digital Services Act protect freedom of expression?' (*Article 19*, 11 February 2021) <https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/> accessed 20 August 2021

B

Balkin J, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011

Barata J, 'Positive Intent Protections: Incorporating a Good Samaritan principle in the EU Digital Services Act' (*Center for Democracy & Technology*, 29 July 2020) <https://cdt.org/wp-content/uploads/2020/07/2020-07-29-Positive-Intent-Protections-Good-Samaritan-principle-EU-Digital-Services-Act-FINAL.pdf> accessed 20 August 2021

Bayer J, *Between Anarchy and Censorship – Public discourse and the duties of social media* (CEPS 2019)

Berthélémy C and Penfrat J, 'Platform Regulation Done Right - EDRi Position Paper on the EU Digital Services Act' (*EDRi*, 9 April 2020) <https://edri.org/wp-content/uploads/2020/04/DSA_EDRiPositionPaper.pdf> accessed 20 August 2021

Bertolini A, Episcopo F and Cherciu N, *Liability of online platforms* (European Parliament 2021)

Bloch-Wehba H, 'Global Platform Governance: Private Power in the Shadow of the State' (2019) 72 Smu L. Rev. 27

Bloch-Wehba H, 'Automation in Moderation' (2020) 53 Cornell Int'l LJ 41

Borgesius FJZ, 'Strengthening legal protection against discrimination by algorithms and artificial intelligence' (2020) 24(10) The International Journal of Human Rights 1572

C

Cambridge Consultants, ‘Use of AI in online content moderation - Report produced on behalf of OFCOM’ (*OFCOM*, 18 July 2019) <https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf> accessed 20 August 2021

Castets-Renard C, 'Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement' (2020) U Ill JL Tech & Pol'y 283

Cauffman C and Goanta C, ‘A New Order: The Digital Services Act and Consumer Protection’ (2021) 00 European Journal of Risk Regulation 1

Cobbe J, ‘Algorithmic Censorship by Social Platforms: Power and Resistance’ (2020) Philos. Technol 1 <https://doi.org/10.1007/s13347-020-00429-0> accessed 20 August 2021

Council of Europe, ‘Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems’ (8 April 2020) <https://rm.coe.int/09000016809e1154> accessed 20 August 2021

Council of the European Union, ‘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 - Presidency compromise text on Chapters II, IV and V, with respective recitals’ 9288/1/21 REV (16 June 2021) <https://data.consilium.europa.eu/doc/document/ST-9288-2021-REV-1/en/pdf> accessed 20 August 2021

D

De Gregorio G, ‘Democratising online content moderation: A constitutional framework’ (2020) 36 Computer Law & Security Review 1

De Streel A and others, *Online Platforms' Moderation of Illegal Content Online* (European Parliament 2020)

Department for Digital, Culture, Media & Sport and Home Office, ‘The government report on transparency reporting in relation to online harms’ (*GOV.UK*, 15 December 2020) <https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/government-transparency-report> accessed 20 August 2021

Directorate-General for Justice and Consumers, ‘Countering illegal hate speech online - 5th evaluation of the Code of Conduct’ (*European Commission*, June 2020)

<https://ec.europa.eu/info/sites/default/files/codeofconduct_2020_factsheet_12.pdf> accessed 20 August 2021

E

EFF, 'Preserve What Works, Fix What is Broken: EFF's Policy Principles for the Digital Services Act' (*EFF*, 2020) <https://www.eff.org/files/consolidatedeupolicyprinciples.pdf> accessed 20 August 2021

Engler A, 'Platform data access is a lynchpin of the EU's Digital Services Act' (*Brookings*, 15 January 2021) <https://www.brookings.edu/blog/techtank/2021/01/15/platform-data-access-is-a-lynchpin-of-the-eus-digital-services-act/> accessed 20 August 2021

European Commission, 'Code of Conduct on countering illegal hate speech online' (*European Commission*, 30 June 2016) <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 20 August 2021

European Commission, 'Tackling Illegal Content Online. Towards an enhanced responsibility of online platforms' (Communication) COM(2017) 555 final

European Commission, 'Shaping Europe's digital future' (Communication) COM(2020) 67 final 1

European Parliament, IMCO Committee, 'Draft report 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' PE693.594v01-00 (28 May 2021) <https://www.europarl.europa.eu/doceo/document/IMCO-PR-693594_EN.pdf> accessed 20 August 2021

European Parliament, LIBE Committee, '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' PE692.898v06-00 (28 July 2021) <https://www.europarl.europa.eu/doceo/document/LIBE-AD-692898_EN.pdf> accessed 20 August 2021

F

Facebook, 'Community Standards Enforcement Report' (*Facebook*, 2021) <https://transparency.fb.com/data/community-standards-enforcement> accessed 20 August 2021

Frosio G, 'The Death of 'No Monitoring Obligations': A Story of Untameable Monsters' (2017) 8 JIPITEC 199 <https://www.jipitec.eu/issues/jipitec-8-3-2017/4621/JIPITEC_8_3_2017_199_Frosio> accessed 20 August 2021

Frosio G, 'Algorithmic Enforcement Online' in Torremans PLC (ed), *Intellectual Property and Human Rights* (Kluwer Law International 2020)

Frosio G and Geiger C, 'Taking Fundamental Rights Seriously in the DSA's Platform Liability Regime' (2020) European Law Journal (forthcoming) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3747756> accessed 20 August 2021

G

Gillespie T, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media* (Yale University Press 2018)

Gillespie T, 'Content moderation, AI, and the question of scale' (2020) 7(2) Big Data & Society 1 <https://journals.sagepub.com/doi/full/10.1177/2053951720943234> accessed 20 August 2021

Google, 'YouTube Community Guidelines enforcement' (*Google*, 2021) <https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB> accessed 20 August 2021

Gorwa R, Binns R and Katzenbach C, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7(1) Big Data & Society 1 <https://journals.sagepub.com/doi/full/10.1177/2053951719897945> accessed 20 August 2021

H

Hern A, 'Online safety bill 'a recipe for censorship', say campaigners' (*The Guardian*, 12 May 2021) <https://www.theguardian.com/media/2021/may/12/uk-to-require-social-media-to-protect-democratically-important-content> accessed 20 August 2021

K

Koene A and others, *A governance framework for algorithmic accountability and transparency* (European Parliamentary Research Service 2019)

Komaitis K, 'The Digital Services Act is tiptoeing towards regulatory failure' (*Open Access Government*, 7 June 2021) <https://www.openaccessgovernment.org/digital-services-act/112311/> accessed 20 August 2021

Kroll J, ‘Accountable Algorithms (A Provocation)’ (*Media@LSE*, 10 February 2016) <https://blogs.lse.ac.uk/medialse/2016/02/10/accountable-algorithms-a-provocation/> accessed 20 August 2021

Kuczerawy A, ‘The Good Samaritan that wasn’t: voluntary monitoring under the (draft) Digital Services Act’ (*Verfassungsblog*, 12 January 2021) <https://verfassungsblog.de/good-samaritan-dsa/> accessed 20 August 2021

L

Land MK, ‘Regulating Private Harms Online: Content Regulation under Human Rights Law’ in Braman S (ed), *Human Rights in the Age of Platforms* (MIT Press 2019)

Llansó EJ, ‘No amount of “AI” in content moderation will solve filtering’s prior-restraint problem’ (2020) Big Data & Society 1 <https://journals.sagepub.com/doi/full/10.1177/2053951720920686> accessed 20 August 2021

Llansó E, van Hoboken J and Harambam J, ‘Artificial Intelligence, Content Moderation, and Freedom of Expression’ (2020) Transatlantic Working Group on Content Moderation Online and Freedom of Expression <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 20 August 2021

M

‘Manila Principles on Intermediary Liability’ (2015) <https://manilaprinciples.org/> accessed 20 August 2021

Minister of State for Digital and Culture, *Draft Online Safety Bill* (CP 405, 2021)

Mostert F, ‘‘Digital due process’: a need for online justice’ (2020) 15(5) Journal of Intellectual Property Law & Practice 378

Mostert F and Urbelis A, ‘Social media platforms must abandon algorithmic secrecy’ (*Financial Times*, 17 June 2021) <https://www.ft.com/content/39d69f80-5266-4e22-965f-efbc19d2e776> accessed 20 August 2021

Murphy H, ‘Facebook’s Oversight Board: an imperfect solution to a complex problem’ (*Financial Times*, 17 May 2021) <https://www.ft.com/content/802ae18c-af43-437b-ae70-12a87c838571> accessed 20 August 2021

N

Nahmias Y and Perel M, 'The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations' (2021) 58 Harv J on Legis 145

P

Pasquale F, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2016)

Perel M and Elkin-Koren N, 'Accountability in algorithmic copyright enforcement' (2016) 19 Stanford Technology Law Review 473

Pollicino O, 'Digital Private Powers Exercising Public Functions: The Constitutional Paradox in the Digital Age and its Possible Solutions' (*European Court of Human Rights*, 2021) (early draft) <https://echr.coe.int/Documents/Intervention_20210415_Pollicino_Rule_of_Law_ENG.pdf> accessed 20 August 2021

Q

Quintais JP and Schwemer SF, 'The Interplay between the Digital Services Act and Sector Regulation: How Special is Copyright?' (*SSRN*, 10 May 2021) (draft) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3841606> accessed 20 August 2021

Quintel T and Ullrich C, 'Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond', in Petkova B and Ojanen T (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar 2019)

S

Sander B, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation' (2020) 43 Fordham Int'l LJ 939

'The Santa Clara Principles on Transparency and Accountability in Content Moderation' (2018) <https://santaclaraprinciples.org/> accessed 20 August 2021

Sartor G and Loreggia A, *The impact of algorithms for content filtering or moderation - "Upload filters"* (European Parliament 2020)

Schmon C, 'UK's Draft Online Safety Bill Raises Serious Concerns Around Freedom of Expression' (*EFF*, 14 July 2021) <https://www.eff.org/it/deeplinks/2021/07/uks-draft-online-safety-bill-raises-serious-concerns-around-freedom-expression> accessed 20 August 2021

Secretary of State for Digital, Culture, Media & Sport and Secretary of State for the Home Department, *Online Harms White Paper* (CP 57, 2019)

Sénat, 'Censure de la loi AVIA: il faut combattre la haine sur internet sans fragiliser la liberté d'expression' (Sénat, 19 June 2020) <https://www.senat.fr/presse/cp20200619b.html> accessed 20 August 2021

Senftleben M and Angelopoulos C, 'The Odyssey of the Prohibition on General Monitoring Obligations on the Way to the Digital Services Act: Between Article 15 of the E-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market' (*SSRN*, 22 October 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3717022> accessed 20 August 2021

Sharpe R, 'Letter: Big Tech auditors need access to algorithms' (*Financial Times*, 23 June 2021) <https://www.ft.com/content/a3fcaf74-4900-44fd-852a-74727a67e973> accessed 20 August 2021

Shattock E, 'Self-regulation 2:0? A critical reflection of the European fight against disinformation' (2021) Harvard Kennedy School Misinformation Review <https://misinforeview.hks.harvard.edu/wp-content/uploads/2021/05/shattock_self_regulation_european_disinformation_20210531.pdf> accessed 20 August 2021

Smith G, 'Harm Version 3.0: the draft Online Safety Bill' (*Informm*, 1 June 2021) <https://inforrm.org/2021/06/01/harm-version-3-0-the-draft-online-safety-bill-graham-smith/#more-49278> accessed 20 August 2021

Smith M, *Enforcement and cooperation between Member States – E-Commerce and the future Digital Services Act* (European Parliament 2020)

Stacey K and Murphy H, 'Now Republicans and Democrats alike want to rein in Big Tech' (*Financial Times*, 12 January 2021) <https://www.ft.com/content/e7c1a64f-b2d9-423b-a86c-f36d1c4e71b7> accessed 20 August 2021

Suzor N, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4(3) Social Media and Society 1 <https://journals.sagepub.com/doi/10.1177/2056305118787812> accessed 20 August 2021

T

Tambini D, 'Media Policy in 2021. As the EU takes on the tech giants, will the UK?' (*Media@LSE*, 12 January 2021) <https://blogs.lse.ac.uk/medialse/2021/01/12/media-policy-in-2021-as-the-eu-takes-on-the-tech-giants-will-the-uk/> accessed 20 August 2021

U

UNGA, 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (29 August 2018) UN Doc A/73/348

UNGA, 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (9 October 2019) UN Doc A/74/486

V

Vermeulen M, 'The Keys to the Kingdom. Overcoming GDPR concerns to unlock access to platform data for independent researchers' (*Knight First Amendment Institute*, 27 July 2021) <https://knightcolumbia.org/content/the-keys-to-the-kingdom> accessed 20 August 2021

W

Windwehr S and Schmon C, 'Our EU Policy Principles: Procedural Justice' (*EFF*, 27 July 2020) <https://www.eff.org/deeplinks/2020/07/our-eu-policy-principles-procedural-justice> accessed 20 August 2021

Wischmeyer T, 'What is illegal offline is also illegal online: the German Network Enforcement Act 2017' in Petkova B and Ojanen T (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar 2019)

Woods L, 'Overview of Digital Services Act' (*EU Law Analysis*, 16 December 2020) <http://eulawanalysis.blogspot.com/2020/12/overview-of-digital-services-act.html> accessed 20 August 2021

Y

Yeung K, 'Responsibility and AI: A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework (Council of Europe study DGI(2019)05)' (*Council of Europe*, 2019) <https://rm.coe.int/responsability-and-ai-en/168097d9c5> accessed 20 August 2021

Appendix

THE AEQUITAS PRINCIPLES ON ONLINE DUE PROCESS[191]

[Relevant excerpts]

[…]

3. Platform Decision-Making Concerning Online Criminal Content – Measures to Address Under-blocking

Platform decision-making concerning criminal content should include the following platform review and judicial review processes.

Platforms must comply with court and agency orders that require take-down of criminal content. Additionally and to the extent permitted by law, platforms should also respond to other notifications of criminal content. For these notifications, platforms should establish prompt, transparent and efficient procedures for submitting and reviewing criminal content notifications and for implementing effective remedies.

Unless provided for by law, platforms should prescribe the content and form of a proper notification together with:

1. Applicable deadlines
2. Timelines for follow-up by platforms
3. An appropriate format for the submission of evidence, and
4. A procedure for the removal of criminal content in a timely manner or, in some instances where warranted, on an expedited basis.

Platform decision-making concerning the removal of criminal content may include or provide for the following platform review and judicial review processes.

1. In urgent circumstances, platforms should make provision for the expedited removal of (serious/flagrant/egregious) criminal content or content which involves the (serious/flagrant/egregious) violation of human rights and which cause imminent harm.

[191] See n84.

Platforms should remove or disable access to such content (as soon as is practically possible and on an urgent basis) following receipt and review of the notification.

2. Expedited removal should follow an expedited review by the platform when so required by law, such as pursuant to court orders and/or notifications by law enforcement authorities, or when requested by trusted parties, such as other relevant government agencies or well-established relevant non-governmental organisations (NGOs).
3. In other circumstances, platforms should make provision for the removal of criminal content or content which involves the violation of human rights in a timely manner. Platforms should remove or disable access to such content as soon as is reasonably possible following receipt and review of the removal order or notification.
4. Due process must be maintained also at the source; therefore, prior to issuing a content removal notification, notifiers such as law enforcement authorities, government agencies and NGOs should establish that they have the appropriate legal basis, such as probable cause or reasonable grounds, to conclude that the activity being notified is illegal.
5. Platforms should limit or remove the ability to use expedited and other procedures in cases of notifiers who have consistently high false positive rates over a specified period of time and/or misuse the expedited procedures to suppress lawful expression. Cases when take-downs are mandated by law or court order and/or notification by a law enforcement agency should be excepted from this principle.

4. Platform Decision-Making Concerning User-Generated Content – Measures to Address Over-blocking

4.1 Overview

Platform decision-making concerning user-generated content should include the following platform review and judicial review processes.

Platforms must comply with laws and court and agency orders that require take-down of illegal content, such as copyright-infringing content. Additionally and to the extent permitted by law, platforms may moderate user-generated content either in accordance with their own standards or based on take-down complaints or take-down notifications ("complaints)". In these circumstances, platforms should provide for prompt, transparent and efficient review of complaints that users submit against take-downs of their content, and implement effective remedies, including content restoration (put-back).

Platforms should establish a clear, simple and easy-to-understand procedure for notifying users about take-down actions or complaints that the platforms receive. Platforms should prescribe the content of a proper complaint and counter-notice and establish procedures to handle unjustified take-down complaints.

Procedural matters prescribed by platforms should include:

1. Complaint and counter-notice deadlines
2. Timely follow-up actions by platforms
3. An opportunity for a complainant to submit evidence in the appropriate format, as specified by the platform, and within a reasonable time frame, and
4. Appropriate and quick appeals procedures, including an opportunity for the parties to be heard, resulting in a timely review decision.

Platforms should not make it unduly difficult to provide evidence, nor should a platform establish stricter requirements than those envisaged by law.

Platform decision-making concerning the removal of user-generated content may include or provide for the following platform review and judicial review processes.

1. Even when they are not required to do so by law, platforms should, where appropriate, establish a notice-and-take-down procedure which allows for the timely removal of law-violating content or content which involves the violation of a platform's standards. Platforms should provide for prompt, transparent and efficient review of complaints. Where appropriate, platforms should remove or disable access to such content (as soon as is reasonably possible) following receipt and review of the complaint. In appropriate circumstances, platforms should also provide for other effective remedies including but not limited to: notice-and-notice, content-labelling, warning, stay-down, suspension, counterspeech, account termination, de-indexing, unmasking, blocklisting and assigning strikes as an alternative to take-downs.
2. Platforms should notify users whose content is removed about the specific reasons for the removal. Platforms should also inform users, in a simple and clear manner, of procedures such as counter-notices which are available to appeal the removal of their content. Such procedures should be easy to understand, user-friendly and outlined in simple terms.

3. Platforms' and complainants' decision-making concerning user-generated content shall not, in principle, prevent users from making content available which is lawful. This includes content covered by freedom of expression, fair use, and other exceptions or limitations to the rights of others, as long as such content does not violate a platform's standards.
4. In order to avoid over-blocking and preserve human rights and fundamental freedoms, including freedom of expression and the freedom to conduct a business, governments should create the following legal environment. With the exception of (serious/flagrant/egregious) illegal or (serious/flagrant/egregious) infringing content, platforms should not be liable for keeping content online and available while an assessment of its legality is completed, unless applicable national law requires that platforms provisionally remove the allegedly infringing material before such assessment is completed.

4.2 Appeals Panel

Where a user or complainant files an appeal against a platform's decision to remove or keep available content, the decision should be reviewed by a competent, independent and impartial decision-maker, such as an appeals panel, in a timely manner. On appeal, the appeals panel should assess the compliance of the decision with the platform's standards. Appeals panel members should have adequate legal and professional training, accreditation and independent standing.

The law may require a review by a competent alternative dispute resolution panel, administrative authority or a court – either in all cases or where the party or parties are not satisfied with the decision of the appeals panel. Even when the law does not so require, in cases when the party or parties are not satisfied with the decision of the appeals panel, platforms should facilitate parties' access to appeal to a competent alternative dispute resolution panel, an administrative authority or a court.

4.3 Automated Review

In response to the volume and velocity of complaints for removal, platforms may employ automated review as the first line of content review. Platforms may use automated removals, algorithmic take-downs and big data analytics to cope with the high volume of complaints.

However, where automated reviews and take-downs do not adequately process urgent or complex cases or where disputes arise which require human review, platforms should establish a clear and simple procedure for complainants to request human review.

4.4 Human Review

In cases which require more urgent review or involve complex issues, complaints for removal and appeals of removal decisions should be forwarded to a platform's human review team in a timely manner. Such cases should include those related to the appropriate weighing of rights and interests, in particular human rights and fundamental freedoms, and those where disputes have arisen over automated content removal decisions. The human review team should assess whether the removal has been correctly carried out according to the platforms' standards.

A clear and simple procedure should be made easily available for a timely appeal to an independent appeals panel in those cases where disputes have not been settled despite the review by a platform's human review team.

Human review team members and appeals panel members should be sufficiently trained, prepared and supported in their professional functions, particularly in cases when they are obliged to review distressing content.

4.5 Algorithmic Enforcement

Governments should carefully weigh the benefits and drawbacks of algorithmic enforcement of rights when deciding whether and to what extent the use of algorithmic enforcement tools should be mandated. Accountability, transparency, and non-discrimination requirements should apply to platforms which employ such tools, whether to comply with the law or voluntarily, including:

1. Algorithm, training data and decision-making logic transparency
2. The application of human-in-command principles
3. External audits
4. Liability and redress for AI-generated harm, and
5. Transparency reports.

4.6 Full Disclosure and Transparency Principles

Platforms should provide clear information to users about the platforms' standards and procedures. The criteria that the platforms use for removing content or for blocking content should be made clear in a transparent manner. Where a user is affected by a platform decision, platforms should inform users about the reasons for the decision. This should include which law and/or platform

standards the content violated and how the manual or automated processes were used to identify the violation.

Platforms should also publish general information about their content moderation practices to enable regulators, government bodies, NGOs and other stakeholders to understand these processes and hold platforms and complainants accountable.

Additionally, in the interest of transparency, full disclosure by platforms should include regular publication of the following information:

1. Posts removed
2. Accounts banned / accounts suspended, and
3. The nature of the complaints made.