Who Moderates the Moderators?: A Brief Overlook into How the Digital Services Act Impacts the Economics of Content Moderation Against "Over-Blocking"

Martim Farinha and Diogo Brandão

Abstract This chapter tackles the intricacies surrounding the subject of content moderation in online platforms, concerning the problem of "over-blocking", the chilling effects it causes on the freedom of speech of users and how it affects user empowerment. It does so by analysing the evolution of the legal framework on content moderation of the last twenty years, focusing the Atlantic dialogue between European Union Law, the e-Commerce Directive, and the Content Decency Act and Digital Millennium Copyright Act. Concretely, it demonstrates the previous framework's pitfalls in ensuring the fundamental rights of freedom of expression and right to information, and how the economics of content moderation allowed for a "false positives" phenomenon ("over-blocking"). It then argues that the more pernicious effect of this comes in the form of chilling effects on users being weaponized for commercial and or political means. It concludes by scrutinizing the recent Digital Services Act's provisions regarding the viability of the procedural mechanisms of redress for users and the accountability of the platforms and third-party stakeholders for misuse of content moderation, whether through algorithmic or individual means. The chapter resorts to a methodology of doctrinal research focused on primary and secondary sources of law.

Keywords Content Moderation · Fundamental Rights · Digital Services Act · Law and Economics · Over-blocking

M. Farinha (✉) · D. Brandão
Nova School of Law, Lisbon, Portugal
e-mail: martim.farinha@novalaw.unl.pt

D. Brandão
e-mail: diogo.brandao@novalaw.unl.pt

G. Enes et al. (eds.), *A Digital Europe for Citizens*,
https://doi.org/10.1007/978-3-032-02500-5_12

1 Introduction

Recent years have been marked by an enhanced public perception of the proliferation of illegal content online (from violations of intellectual property rights (IPR), to child pornography, hate speech, terrorist propaganda, misinformation, etc.) and its undesirable impact[1] in society. This has prompted a wider reflection across political lines, on the roles of online content-sharing platforms, their practices on content moderation and their responsibilities on the harms caused by their (in)action.

While there are several possible business models, most online content sharing platforms adopted a "freemium model", deriving most of their profits (Desjardins 2019) from the activities connected to the processing of personal data of their users, which allows them to create detailed profiles for personalized targeted advertising and to tailor their recommender systems, to continuously promote user activity in the service. It is understood that while these platforms act as intermediaries, assuming a passive stance towards user content unlike traditional publishers and editors, they still derive value from this content, which may not always be lawful. They may not promote such illegal activities, or are even fully aware of them, they nevertheless benefit from them.

Whether the damages from these activities are caused to singular private parties (rightsholders, victims of defamation, etc.), groups or to society as a whole (hate speech, child pornography, terrorism propaganda), tackling illegal content online has been framed as a matter of determining the most optimal model for the attribution of liability or safe harbour to these platforms (designing the exemption of liability or the liability itself may yield similar results in practice) (Nordemann 2018; Husovec 2017).

Determining the appropriate liability rule for these online hosting platforms requires an analysis of the incentives to detect and remove illegal content. A liability rule should be designed to minimize the combined costs of harm and of detection and removal, considering both the private costs and social costs. This rule should also consider the role of each actor in this matter, since there may be other parties than platforms, such as users of third party-victims or public bodies, that may have better information available to them to detect illegal content and its harm. Therefore, it is a difficult exercise to determine which policy is most proportionate and effective.

This chapter purports to provide a brief overlook on the way in which these concerns have been tackled by the European Union's legislature and judges and on the role the recent Digital Services Act (DSA)[2] plays in mitigating the "over-blocking" phenomenon. The analysis is conducted under the lens of user empowerment, with each section demonstrating how each specific aspect impacts the control users have over their own content and the mechanisms they have to react to content moderation. Additionally, given the horizontal character of the DSA as an asymmetric regulation,

[1] An example of this pertains to Molly Russel's suicide in 2017, following exposure to graphic content in Instagram. See < https://www.bbc.com/news/uk-england-london-54307976 > .

[2] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

with a wide subjective and objective scope of application, the analysis will be focused on how effective its provisions will be when applied to the very large online platforms (VLOP) and very large search engines (VLSE).[3]

The second section of this chapter provides a legal context of relevant policies and court decisions on content moderation to highlight the European Union's (EU) key motivations and concerns regarding the phenomenon under analysis.

The third section of the chapter analyzes the "over-blocking" phenomenon at a conceptual level and assesses how it has been fueled by platforms growing use of algorithms. Moreover, this part of the chapter describes the chilling effect that has concerned scholars and policy makers over recent years and frames some of the key concerns that the EU legislator purported to mitigate through the DSA. These considerations are fueled by practical examples of online platforms' content moderation systems, particularly that of Youtube.

In turn, the fourth section of the chapter offers a brief examination of the economics of content moderation law and provides palpable outcomes and behaviors stemming from the prior regulation on the topic. It does so by providing both a user and platform perspective.

The fifth and final section of the chapter focuses on the DSA per se and assesses its prospective merits in combatting the chilling effect of the "over-blocking" phenomenon. It specifically emphasizes the transparency obligations imposed by the DSA regarding action and notice and redress mechanisms, while also focusing on the functioning of the mechanisms per se. The analysis is also supported by occasional analogies with the current EU legislation on personal data protection, especially regarding the role of regulators, given the General Data Protection Regulation's (GDPR) connection to the DSA's goals and its standing in the Digital Single Market.

The research conducted is theoretical and doctrinal in nature, having been conducted through the analysis of primary and secondary legal sources in the form of scholar articles, court decisions, policies and other legal documents.

2 Brief Legal Context of Content Moderation: The Road until Now

It is relevant to assess the way legislators have tackled the topic of content moderation to give a comprehensive notion of the phenomenon under discussion and to justify the need for tackling it. For purposes of this section, emphasis is to be given to an EU

[3] The DSA is a "layered" or "pyramidal" regulation: as the "size" of the service provider increases, so do the sets of obligations applicable. At the very top, the VLOP and VLSE have to comply to the entirety of the DSA. G'sell, Florence (2023) The Digital Services Act: a General Assessment. In: Antje von Ungern-Sternberg (ed.), Content Regulation in the European Union—The Digital Services Act, Schriften Des Irdt—Trier Studies On Digital Law, Vol. 1, Verein für Recht und Digitalisierung e.V., Institute for Digital Law (IRDT), Trier April 2023, available on SSRN, p.5.

context, focusing particularly on the E-Commerce Directive (ECD)[4] and Directive 2019/790 (DSMD)[5] from a policy perspective, as well as on relevant European Court of Justice (CJEU) jurisprudence.

2.1 Policy

From the late 90s, lawmakers on both sides of the Atlantic established the principle of "safe harbour" for online intermediary service providers (OSP), to protect these entities from secondary liability that could arise due to the unlawful activities of their users, namely the hosting of unlawful content.

These policy decisions were justified on several concerns, namely:

1. The common understanding that if these OSP performed a passive role of transmission, caching and hosting of content, they should not be directly responsible for it.
2. The notion that due to the anonymous, probable insolvent and jurisdictionally unreachable character of the actual primary infringer, OSP would be the preferred target of lawsuits by third parties seeking reparations. Intermediaries act as "choke points", where this strategy allows claimants to secure relief in a single procedure against multiple infringements from several perpetrators.
3. The understanding that, without these "shields" shifting some of the liability, these activities would be so risky as to, either resulting in over-enforcement (also known as "over-blocking") by overtly cautious intermediaries, silencing the legitimate users (Dinwoodie 2017); or
4. The conclusion that, when considering the factors mentioned above, the provision of these services risks becoming commercially unviable or overburdened, stunting the development of information society services and overall technological progress. These fears of stagnation could culminate in the inability of American/European companies to be competitive in global markets.

Considering these risks, many legislators adopted provisions in those early days of the Internet, that either established a "positive" (circumstances that establish liability) or "negative" (actions and standards that immunize intermediaries against claims) definitions of safe harbour (Dinwoodie 2017).

[4] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce').

[5] 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.

In the EU this was established with the ECD,[6] which defined the conditions intermediaries that provide online services required for an horizontal liability immunity regarding the goods, services, and contents they provide.[7] For hosting services, the requirements hinged on the service not having actual knowledge of the illegal activity, and upon obtaining such awareness, acting expeditiously to remove or disable access to the infringing information.[8] The ECD also enshrined the principal of "no general obligation to monitor", which would preclude Member-States from creating such obligations.[9] In summary, the ECD's liability regime requires intermediaries to assume a liability stance regarding their content and to also commit to eliminating illegal contents, as well as the access to said contents upon notification.

However, the ECD displays a visible difficulty in accompanying the evolution of tech services as well as in harmonising a legal framework between the EU Member-States in what concerns the monitoring and supervision of digital service providers' activity, as well as combatting data asymmetries. These difficulties are worsened by the power imbalances between platforms and users, which in turn stem from a lack of transparency over how user data is processed and monitored. Indeed, despite the guarantees created by the e-Commerce Directive, studies conducted by BEUC indicated that about two thirds of a significant number of digitally acquired products did not display proper safety guarantees regarding the monitoring and notification of their contents.[10] This situation became more severe during the Covid-19 pandemic, in which the illegal acquisition of digital goods and services registered a significant increase, in line with growing dissatisfaction with platforms' reporting mechanisms for their content.

At the time, these conditions were considered an open invitation to self-regulation due to the belief that the removal of strict obligations to monitor—while leaving some room for liability -would incentivize hosting services to impose some measure of control upon the content which they receive (Hornik and Villa Llera 2017) The framework also promoted the implementation of "notice-and-takedown" mechanisms for third parties to notify OSP and complaint-and-redress schemes for users whose content had been taken down. Public consultations conducted by the European Commission further demonstrated a general predisposition from market players towards creating simple, transparent, and harmonized standards for purposes of reporting illegal content in the context of the Digital Single Market.[11]

---

[6] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce').

[7] Arts. 12–15 ECD.

[8] Art. 14 ECD.

[9] Art. 15 ECD.

[10] BEUC (2020) "Two-thirds of 250 products bought from online marketplaces fail safety tests, consumer groups find", BEUC Press Release, < https://www.beuc.eu/press-releases/two-thirds-250-products-bought-online-marketplaces-fail-safety-tests-consumer-groups > .

[11] This public consultation, which preceded the proposal for a Digital Services Act, collected the answers of two hundred and eleven companies (ranging from conglomerates to "start-ups") that provide digital services, as well as one hundred and fifty-nine Non-Governmental Organizations

2.2 A Joint Policy and Jurisprudential Analysis

Through the years, the CJEU was "called" several times to interpret these provisions (Quintais et al. 2022), namely the character of an "active" or "passive" host and the scope and matter of the prohibition of general monitoring obligations, in relation to several national laws in different fields, from the protection of intellectual property rights (IPR) to defamation. On the latter, the case law focused itself on whether certain provisions or injunctions concerned general or specific monitoring obligations, where the threshold was located (from a single copy of specific unlawful file on a URL, to all copies of that file, subsequent reposting of copies, variations, etc.). Through the years, the CJEU failed to provide a precise criterion to distinguish permissible specificity from prohibited generality, rather deciding each individual case on proportionality assessments, weighting primarily the impact on fundamental rights and social values (Sartor and Loreggia 2020).

This is most noticeable on various copyright cases, such as C-275/06 Promusicae,[12] C-275/06 Scarlet Extended,[13] and C-360/10 SABAM,[14] in which it was decided that injunctions to filter all copyrighted content were general in scope, infringing on the fundamental rights of users (Art. 8 and 11 of the EU Charter of Fundamental Rights), and thus inadmissible. *Au contraire*, in a defamation case, C-18/18 Glawischnig-Piesczek,[15] the CJEU allowed an injunction to block and prevent the reposting of equal or equivalent content, since the host would not be required to carry out independent assessments and the measured was required to protect the interests at stake (Sartor and Loreggia 2020).

The CJEU found that video-sharing and file-hosting platforms could not be directly liable in C-682/18 and C-683/18 (Youtube, Cyando),[16] and that (in relation to secondary liability) would only have actual knowledge if properly notified of a specific illegal act, "where the application of an exception is not automatically precluded". As many pointed out, on the matter of the right of communication to the public of EU copyright law, the CJEU slowly eroded the paradigm of traditional strict liability, by incorporating elements typical of intermediary, secondary or accessory liability (Quintais and Schwemer 2022).

While the CJEU was "carefully" building its case-law on the interpretation of the Arts. 14 and 15 ECD and Art. 3 of the InfoSoc Directive,[17] the Commission began

(NGOs) and fifty-nine public authorities. European Commission (2020) Summary Report on the open public consultation on the Digital Services Act Package. In < https://digital-strategy.ec.europa.eu/en/library/summary-report-open-public-consultation-digital-services-act-package > .

[12] Judgment of the Court of 29 January 2008, Promusicae, C-275/06, ECLI:EU:C:2008:54.

[13] Judgment of the Court of 24 November 2011, Scarlet Extended, C-275/06, ECLI:EU:C:2011:771.

[14] Judgment of the Court of 16 February 2012, SABAM, C-360/10, ECLI:EU:C:2012:85.

[15] Judgment of the Court of 3 October 2019, Glawischnig-Piesczek, C-18/18, ECLI:EU:C:2019:821.

[16] Judgment of the Court of 22 June 2021, Youtube & Cyando, C-682/18 and C-683/18, CLI:EU:C:2021:503.

[17] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

the process of re-balancing the roles of intermediaries and the problem of tacking illegal content online. This occurred first soft law instruments, such as the 2018 Commission Recommendation,[18] and with hard law proposals such as the DSMD on 2015, and later, with the DSA in 2020 (recently approved in late-2022).

The DSMD and its Art. 17 are the biggest reform in the recent years on platforms' liability (Quintais et al. 2022), concerning IPR. It acts as *lex specialis* of the ECD, by stating that online content-sharing service providers (OCSSPs) carry out acts of communication to the public when they give access to work/subject matter uploaded by their users and that the hosting safe harbour is not applicable (Quintais 2020). OCSSPs may now be held directly liable for their users infringing content, if they cannot prove their best efforts to either:

a) Secure appropriate authorisations from rights holders; or
b) to ensure the (i) unavailability of specific works for which they have been provided with information, (ii) act expeditiously after notices to take down infringing content, and (iii) prevent their future re-upload.

These conditions seem to impose, as it widely criticized in the legislative process, both an upload filtering obligation and a notice-and-stay-down (or re-upload filter) obligation (Quintais 2020). Due to many criticisms from civil society,[19] academia, Non-Governmental Organisations (NGOs)[20] and courts on the matter of upload filters constraining users' fundamental rights of freedom of expression and of accessing information, the DSMD includes some safeguards in this regard.[21]

Platforms must ensure that their measures do not prevent users from uploading their lawful content, including when it relies on exceptions and limitations (E&L), such as parody and quotation, which became mandatory across the EU. Following the Commission's guidance and the AG opinion on C-401/19, under Art. 17, platforms should only filter ex-ante manifestly infringing content, based on a highly probable quantitative assessment (Geiger and Jütte 2021).

Finally, the CJEU's decision on C-401/19[22] cemented the understanding that in Art. 17 of the DSMD there is a clear hierarchy, where the obligation to ensure the availability of lawful content is a obligation of result, that supersedes the (best efforts obligation) to ensure the unavailability of unlawful content (Quintais and Schwemer 2022). It is insufficient for users to merely rely on ex post complaint and redress mechanisms to ensure the accessibility of their user generated content (a

[18] Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online C/2018/1177.

[19] #SaveYourInternet, "The #SaveYourInternet fight against Article 17 [ex Art. 13] continues", at < https://saveyourinternet.eu/ > .

[20] EDRi (2019) Upload filters: history and next steps. in < https://edri.org/our-work/upload-filters-status-of-the-copyright-discussions-and-next-steps/ > , OpenDemocracy (2018) The EU call it copyright, but it is massive Internet censorship and must be stopped. In < https://www.opendemocracy.net/en/can-europe-make-it/civilised-societies-don-t-call-it-censorship-but-copyright/ > .

[21] Art. 17(7)(9) DSMD.

[22] Judgment of the Court of 26 April 2022, Poland v European Parliament and Council of the European Union, C-401/19, ECLI:EU:C:2022:297.

position that was held by France, Spain and Portugal). OCSSPs are required to have ex ante safeguards to avoid "over-blocking" of users' content due to the usage of filtering technologies, "upload filters", which were recognized as implicitly required (Quintais and Schwemer 2022). Only filtering/blocking systems that unequivocally can distinguish lawful from unlawful content without the need for "independent assessment" by OCSSPs are admissible (so, only manifestly infringing content as in *Glawischnig-Piesczek*), without infringing the prohibition of general monitoring obligation under Art. 17(8) DSMD (and the ECD) (Quintais and Schwemer 2022). Users' have (fundamental) rights, not mere defenses.

Still, while both the AG and the CJEU aimed to uphold Art. 17 in balance with the prohibition of general monitoring obligations and the fundamental rights if users, by emphasizing the objective of avoiding "over-blocking" in combination with good redress mechanisms, in practice, some matters were left unresolved. In practice, there is no concrete definition of what acceptable error rates for content filtering are. Many "key" details regarding these "redress" procedures and how they look like in practice were not explored in the decision, nor transparency obligations or compliance instruments for the OCSSPs. Additionally, the directive and the case concerned only copyright infringement, leaving many doubts persisting in relation to other types of illegal content and even "harmful" content,

3 The Chilling Effect of "Over-Blocking"

Despite the context presented above, worries persist and are fueled by various factors—both legal and technical -, out of which this chapter highlights the chilling "over-blocking" effect caused by algorithmic content moderation. The next section will focus on describing the phenomenon and highlighting its pernicious effects for the topic of the chapter. For these purposes, it is relevant to clarify the role algorithms (Hill 2016)[23] play in fueling "over-blocking" and how the chilling effect affects users and content moderation.

3.1 The Role of Algorithms in Content Moderation

Although protected by safe harbors in both side of the Atlantic, many of the most influential content-sharing platforms suffered pressures to self-regulate, leading to major investments in technological measures aimed at proactively detecting and removing illegal content. As these mega-platforms grew (and the controversial incidents multiplied), the need to turn to AI for moderation became undeniable. In turn, the task to

[23] For purposes of the research, we follow the simplified definition of algorithms as a "finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions".

monitor all content uploaded for illegal activity became humanly impossible. This context led to the adoption of algorithms trained to detect illegal content using several techniques, such as matching, hash-matching, perceptual-hashing, classification, and prediction (using natural-language-processing).

In the current state of the art, depending on the type of illegality pursued, each OSP may employ different combinations of these systems to monitor user content (Gorwa et al. 2020). However, even with the major breakthroughs in their capabilities, evidence shows that these systems still have major problems at understanding many caveats of human language, especially concerning context-dependent nuances. A notable example of this is that of Youtube's Content ID which, while highly advanced at creating and detecting quality fingerprints of copyrighted content, is not actually adept at evaluating actual infringements, due to not understanding the applicability of E&L.

Concerns have arisen over the deployment of fully automated moderation resulting in systematic "over-blocking" when applied to more context-dependent judgments, such hate speech or misinformation, the fears that the deployment of fully automated moderation will result in systematic over-blocking become exacerbated. On this note, scholars (Quintais et al. 2022) have highlighted how the European Commission's guidance[24] on OCSSP clarifies that automated filtering and blocking measures should be reserved for what the EC classifies as "manifestly infringing" and "earmarked" content.

Due to these concerns, platforms usually (claim to) include a "human-in-loop" to review the decisions of their algorithms, but these do not necessarily confer neither security nor legitimacy to these systems (Gorwa et al. 2020). Once again, Youtube's Content ID exemplifies this conundrum, since the automated decision are confirmed by rightsholders, which have an interest in concluding on the inapplicability of fair-use standards or E&L (Kaye and Gray 2021). In other cases, the "human" in these procedures lacks the appropriate tools, formation and time to adequately decide on the matters under review. In opposite sides, these "content moderators" may suffer psychological damage from the exposure to violent material (Newton 2019)[25] or help train the automated filtering with bias against certain people, groups, political orientations (Haimson et al. 2021).

Despite a growing number of controversial takedowns decisions and recent empirical studies which report that automated copyright enforcement systems incur in substantial over-blocking of legitimate content (Bar-Ziv & Elkin-Koren 2018; Erickson & Kretschmer 2018; Urban et al. 2017) on the regulatory trend of promoting *ex ante* decisions has persisted, as it was referred in regarding the DSMD.[26]

[24] European Commission (2021) Communication from the Commission to the European Parliament And The Council—Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market, COM/2021/288 final.

[25] https://perma.cc/GD4C-U8AE.

[26] See Sect. 2 of this chapter.

3.2 The Chilling Effect of Takedowns

Empirical studies indicate that takedown notices produce a chilling effect against users. This occurs with both formal notices and through internal alternative dispute resolution (ADR) systems in these online services. After receiving a DMCA notice (for copyright infringement), it has been reported that 66% of users would not even attempt to challenge it, even "when they believed it to be wrong or inaccurate". This position was justified on the concerns regarding "legal costs (81%) and avoiding additional risks (legal or otherwise) (53%) and a lack of legal knowledge (34%)" (Penney 2019).

A similar effect can be observed in Youtube's Content ID system. This automated system not only detects the fingerprint of copyrighted material in users' content, but also informs rightsholders and allows them to either: block the video, monitor the statistics or claim it (the video is monetized, and the ad-revenue reverts to the rightsholder). Users may dispute the claim. Afterwards, rightsholders have 30 days to decide whether to release the claim, uphold it or issue a removal request. If they choose the latter two, they automatically win the dispute, and the uploader receives a strike. If a user receives three strikes, whether legitimate or not, their account (which may be their livelihood) is terminated (Kaye and Gray 2021). Youtube's own transparency report reveals that over-blocking is real and that users are highly discouraged from fighting it (Kaye and Gray 2021).

If a user wishes to learn more on how these systems work, under what rules they are operating under when uploading content, they will quickly be "submerged" when trying to study the T&C regarding content moderation. Platforms have, over time, published longer T&C, more documents, more add-ons (on more URLs), with more (kinds of) rules, which have become more intricate, vague and complex. They have essentially created an ever-updating opaque web of legal documents that the normal user cannot adequately understand (Quintais et al. 2022).

Regular users do not engage with content moderation redress mechanisms when their content is blocked/filtered, even they believe that the decisions were wrong. Instead, they seem to take the notices as warnings, resorting to self-censorship in the future. "Content creators" in Youtube and other OCSSPs, either resort to tactics to trick the systems into not erroneously flagging their content (Kaye and Gray 2021), or just "surrender" and accept the loss of revenue due to the licensing system in Art. 17. The pre-DSA framework does not prevent the chilling effect of "over-blocking". It seems to allow it.

4 The Economics of Platforms' Liability and Content Moderation

To provide a better understanding of the phenomenon under analysis, it is relevant to assess the economics of platforms' liability and content moderation from the standpoint of both the platforms and of the users. This analysis will contextualize some of the goals of the DSA, which will be analyzed in the next section of the chapter.

4.1 The Platforms' Perspective

The platforms themselves have several incentives to monitor and remove illegal content, as they can be negatively affected by its proliferation. Much of this content can result in the deterioration of the user-experience, which can lead to a reduction in customers' participation or activity. OSP risk losing users to more trustworthy and less "toxic" competitors (Buiten et al. 2019). Indeed, the presence of illegal content may cause harm to the reputation and credibility of a platform with its users, customers, and commercial partners. Obscene or highly problematic content can scare off advertisers, who do not wish to see their products placed alongside terrorist propaganda, xenophobic or pornographic videos (Gillespie 2017).

The initiative in implementing content moderation measures may also be driven for the purpose of preventing legislators from enacting new regulations, by convincing them that their self-regulation suffices (Husovec 2017). As their influence in society expands, monitoring may also be motivated out of a sense of "public obligation", or as means of answering to criticisms form activists and journalists (Gillespie 2017). Beyond the purely profit-maximising incentives, these platforms may also be committed to nurturing a healthy and creative online community as a goal by itself (Buiten et al. 2019), or at least, to appear to do so.

In terms of costs, platforms will need to deploy several measures for detection, removal, and sanctioning, which will be affected by different factors. These measures can be automated to varying degrees depending on the concrete case. They will include the development and maintenance of detection and flagging software, as well as notice and takedown systems, with varied degrees of human-reviewer intervention. These will be affected by the size of the platform (larger OSP may benefit from data-drive economies of scale, which will increase efficiency and lower costs), type of harmed party, the business model (freemium vs payment-for-access, social media) and the type of illegal material (automation is less prone to error if the illegality is not context dependent) (Buiten et al. 2019).

4.2 The Users' Perspective

When approaching this matter from a third-party perspective, the costs and incentives differ across the type of victim. In the case of IPR violations, rightsholders may have the means to notify OSP about infringements and may leverage their position so that content is proactively detected and removed. Some platforms, such as Youtube, offer rightsholders a third option besides just accepting or removing: monetising the infringing content in their favor. Victims of hate speech have little means to prevent this kind of harm against the primary infringers and must rely of notifying the platforms (Buiten et al. 2019).

As previously referred, there is also content which harms society, without clear individual victims, such as terrorist propaganda. If no private parties act, unless public law enforcement acts, this material could stay up. Both users and third parties have no means of removing illegal content once detected if platforms do not provide those means. Platforms often lack the relevant information to detect said content, even if motivated to remove it. There is a clear link between the costs of each party in this relationship. Users are a heterogeneous group, sensible to different interests and incentives. Some may just consume content while others produce and share. For the latter, there is a wide range between amateurs and professional creators, motivated by profit, who are also interested in the removal of unlawful and toxic content.

The problem here lies on whether the platforms consider that the costs of implementing said infrastructure (user-friendly notice-and-takedown systems), and acting upon it, do not go beyond the incentives to do so (Buiten et al. 2019). Therefore, a liability regime[27] for OSP is necessary to internalize these externalities. Two solutions should be avoided: full exemption and strict liability, as the latter would inevitably promote excessive over-blocking.

5 The Digital Services Act's Role in Combatting "Over-Blocking"

The context provided above contextualizes some of the key concerns that motivated the drafting and development of the DSA which, alongside the Digital Markets Act (DMA), assumes a key position[28] (Kaiser and Ratcliff 2022) in the European Commission's strategy to tackle data processing by massive online platforms and

[27] European Commission (2017) Working Document on the free flow of data and emerging issues of the European data economy COM(2017) 9 final. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52017SC0002&from=EN.

[28] https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/703347/IPOL_BRI(2022)703347_EN.pdf.

better respond to the challenges of a 'data driven economy'.[29] From the start of its discussions (Morais Carvalho et al. 2021) the DSA (European Commission 2020) focused on revising the liability framework for the actors partaking in the collection, storage, and transmission of data and content in the context of intermediary services (G'sell 2023), aiming to improve the functioning of the Digital Single Market (G'sell 2023).[30] On this note, the DSA's scope encompasses not only service providers that are based in EU Member-States, but also those based in countries outside the EEA and whose activities affect the EU internal market.[31] This greater territorial reach aims to not only protect the users whose data will be processed, but also to keep service providers headquartered in the EU from being at a competitive disadvantage regarding service providers headquartered outside the EU.[32]

The DSA as a piece of legislation, tackles many of its objectives through a "meta-regulation" approach (Zingales 2023), which include several elements of hard and self-regulation. Many provisions establish regulatory principles and obligations that give businesses a significant amount of discretion in their implementation (with standards and codes of conduct) which then are coupled with other obligations for continuous monitoring and evaluation of results, by the businesses themselves (risk assessments, risk mitigation measures, compliance function), authorities (enhanced transparency and access to data by the Digital Coordinators and Commission) and certified third parties (enhanced transparency obligations, independent audits, access to data by vetted researchers). The DSA follows a recent trend of risk-based regulation—or risk-based compliance—(Savin 2022) in European Law, which can also be found in the GDPR (with the Data Protection Impact Assessment and Data Protection Officer as mere examples) and the Artificial Intelligence Act, its proposal and subsequent amendments (Calvi and Kotzinos 2023).

One of the key changes that the DSA promotes is a greater set of responsibilities for intermediary service providers, while maintaining the core principles of exemption of liability for intermediary service providers from the previous e-Commerce Directive. Among these reforms, it includes the "Good Samaritan Principle" and the development of user-friendly notice and take-down mechanism that allow for third parties to notify illegal content. With these mechanisms, third parties must be able to submit detailed notices on the content that they deem to be illegal and incompatible with the service provider's Terms and Conditions (T&C). The service

[29] European Commission (2014) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions towards a thriving data-driven economy. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014DC0442&from=EN.

[30] European Parliament (2020) European Parliament resolution of 20 October 2020 with recommendations to the Commission on the Digital Services Act: Improving the functioning of the Single Market (2020/2018(INL). https://www.europarl.europa.eu/doceo/document/TA-9-2020-0272_EN.pdf.

[31] Art. 1(3) DSA.

[32] These concerns are in line with the Impact Assessment conducted on the DSA itself; European Commission (2020) Impact assessment of the Digital Services Act SWD(2020) 348 final. https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act.

provider will then decide on whether to remove or restrict the notified content, and then inform the notifier of its choice. Equally relevant for purposes of this chapter are the redress mechanisms developed by the DSA regarding content moderation. (Savora 2022). Concretely, the DSA lays down obligations for online platforms to provide users with an internal complaint-handling systems[33] for decisions taken over content. Moreover, it adds a second layer of redress for content moderation by obliging platforms to engage with certified out-of-court dispute settlement bodies,[34] while also requiring transparency enhancing practices from online platforms over how they moderate content.[35]

It is, therefore, clear that the DSA actively purports to empower users by endowing them with the proper tools to handle restrictions on the content they share online, while also creating mechanisms and procedures that aim to prevent OSP from engaging in the "over-blocking" phenomenon. Theoretically, with the DSA comes a framework of greater accountability for OSP and of greater flexibility in terms of how users might exercise their rights over their content. However, despite the legislator's efforts, one of the main reservations highlighted during the DSA's discussion pertained to the pernicious effects of automated decision-making regarding the monitoring of content in digital platforms. The main concern stemmed from how automated decision-making could exacerbate the power asymmetry[36] between users and platforms, rather than mitigate said asymmetries.[37] On this note, scholars (Quintais et al. 2022) have pointed out that automated platforms can unilaterally follow decision patterns and be opaque regarding the factors weighed into consideration. In response to these reservations, some the final revisions[38] on the DSA proposal prior to its approval included endowing users with redress mechanisms for any damages incurred due to platforms' monitoring practices.

Focusing on DSA's wording of service providers' transparency obligations regarding content moderation, on the resources required to implement an effective notice, action, and redress mechanism and on the mechanisms per se, this chapter will aim to assess these measures' purported effectiveness. The assessment to be conducted will also be impacted by the role Digital Services Coordinators (the

[33] Art. 17 DSA.

[34] Art. 18 DSA.

[35] Art 23. DSA.

[36] C-401/19 CJEU para 85 and 86, in which filtering systems that cannot distinguish lawful from unlawful content are incompatible with the right of freedom of expression. See also the thoughts of the Advocate General on this case, concretely 64, 165 and 191 to 193 of his Opinion, in which he tackles the limits of automatic recognition and filtering tools for content moderation purposes.

[37] A notorious example of such an asymmetry pertains to Youtube's content moderation system, based on which users are strongly discouraged from contesting automated decisions over their own content. See http://copyrightblog.kluweriplaw.com/2021/12/09/youtube-copyright-transparency-report-overblocking-is-real/.

[38] European Commission (2021) 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. https://www.europarl.europa.eu/doceo/document/A-9-2021-0356_EN.html?redirect.

primary national authorities designated to ensure the consistent application of this Regulation) might play in the tackling of the "over-blocking" phenomenon. The topic is approached in a manner that clarifies the role played by users, moderators and regulators. Considering that the practical implications of the DSA have yet to be documented, these considerations are purely theoretical.

5.1 On the Effectiveness of the Mechanisms Per Se

This part of the chapter specifically focuses on assessing the practical effectiveness of the DSA's mechanisms for content moderation, considering the context presented thus far, as well as all the considerations presented during the chapter. Specifically, it aims to demonstrate each mechanism's role in promoting user empowerment.

5.1.1 Counter-Notice Mechanisms

Upon assessing the provisions set forth by the DSA regarding redress mechanisms on notice and action, scholars have identified a common "missed opportunity"in the lack of provisions that would entitle users to issue counter-notices (De Streel et al. 2020). An immediate consequence of this is the inherent vulnerability of users, who are put into a more reactive position regarding their rights, seeing how they can only lodge claims following an intermediary's decision.

A strong framework ensures user empowerment by making users able to not only flag content that is illicit, but also to properly defend their own content from automated decisions. This can only be achieved if both service providers and service recipients are able to proactively contribute to creating the intended robust framework. Considering that one of the main concerns that has been pointed out regarding the DSA is the uncertainty of some of its dispositions when put to practice, it is curious that the legislator precluded the opportunity to give users additional motivation in being more active in the regulation of content.

Whether or not the legislator's final decision reflected reservations on internal and external channels being flooded with complaints remains unclear, but the current formulation of the DSA does not properly mitigate these concerns either, as discussed in the previous parts of this section. Thus, from the perspective of empowering the service user, the cost/benefit relation at hand indicates that a counter-notice provision might prove to have been more beneficial in the long run.

5.1.2 Internal Redress Mechanisms

The more recent drafts of the DSA before its approval raised concerns (Zeybek et al. 2022) on how the broadened internal complaint-handling systems risk compromising the functionality of the redress mechanism itself, a statement that merits a

critical assessment. Concretely, Art 20(4) DSA states that platforms should handle complaints in a timely, non-discriminatory, diligent, and non-arbitrary manner, which is a very abstract disposition, regardless of Recital 58 DSA mentioning that systems should promote fair outcomes. The very abstract dispositions of Art. 20 DSA risk incurring in a high degree of unpredictability when it comes to the resolution of complaints, which does not allow for a uniform application of the Regulation's norms, even when admitting different terms and conditions for different service providers.

Therefore, OSP concerns that hosting proper dispute resolution channels might prove to be too costly (Kuczerawy 2022) raises reservations in the sense that these providers can be inclined to make their policies and terms and conditions overly restrictive. Perhaps more concerning than this is the incentive that OSP might have to actively push for external resolution methods that users end up being unable to afford, which can prove most pernicious for purposes of users' interests and their right to redress.

5.1.3 Measures to Tackle Misuse of the Notice Mechanism

Art. 23 of the DSA aims to tackle the problem of "misuse" of the online platform, by recipients of the service that frequently provide manifestly illegal content, and by individuals or entities that frequently submit notices or complaints that are manifestly unfounded—which is at the core of over-blocking. The provision mixes two different opposing issues and seems to solve neither. The article merely establishes an ill-defined bottom floor for both problems, with soft consequences for each. Concretely, it prescribes an obligation to suspend "for a reasonable period of time and after having issued a prior warning", which while it may be welcomed, it can reveal itself insufficient and be considered to constitute a *"loophole"*. By mixing both problems—users uploading illegal content vs. abuse of the notice mechanism—the article actually ensures that the online platforms have enough discretion to treat both problems very differently.

As such, the provision merely establishes that service providers must suspend users that frequently upload manifestly illegal content[39] but does not disallow harsher measures. The service provider may suspend for longer periods or even terminate the provision of the service for uploading content that, while illegal, is not necessarily manifestly illegal. The standard for how frequently those violations need to occur may also "play" against the recipient. Additionally, it is also unclear how the "prior warning" system can be implemented in practice for instances of simultaneous detection of several violations.

On the other hand, service provides may adopt a higher standard to consider a notice "manifestly unfounded" for the purposes of applying this article, allowing the submission of notices and claims that are "just" incorrect or "partly" unfounded. Much content could be unlawfully blocked if doubts arise about its legality (see the problems with exceptions and limitations to copyright protection previously alluded

[39] Fn 68 (2022).

to). Additionally, while well intentioned, Art. 23(3)(b) could also be used to temper with the system in favor of whoever is abusing the notice mechanism: if enough correct notices are made, the mechanism that punishes for misusing the system is not triggered.

Nevertheless, Art. 23(3)(d) is a positive inclusion. If applied correctly, this provision could be a game-changer. Individuals or entities which abuse the notice and claim system with an agenda would be suspended thanks to this criterion. Nevertheless, some skepticism can persist, given that while some may see a pattern of abuse with a clear intention (harassment, censorship, whether economic, political, or ideological), others could conclude that such a malicious intention does not exist. Therefore, a joint application of Art. 23(4) with Art. 14 DSA will be crucial in cases in which there is uncertainty regarding the practical implications of many provisions of the DSA, particularly on how intermediary service providers will comply with them.[40] This is yet another measure that aims to tackle the opaqueness of online platforms by prescribing transparency obligations via the inclusion of clear and detailed policies regarding "misuse" in the terms and conditions. This obligation of transparency is further enhanced with the obligatory inclusion of examples of facts and circumstances that will be considered in the assessment of misuse, providing much necessary legal certainty to users and third parties.

As a conclusive note, this system needs to receive the appropriate attention by OSP or it is bound to fail in its objectives. Many notice and complaint systems are notorious for being abused against users due to lax consequences for misuse and or lack of oversight and enforcement (Appelman 2023). If faux complaints are submitted are followed through, users can be the subject of harsh penalties and damages, while the "complainant" can only be temporarily suspended. One may find this sort of imbalance in the similar issue in the practice of SLAAP lawsuits.

5.1.4 The Out-of-Court Dispute Settlement Mechanism

Out-of-court dispute settling bodies are to be designated by the Digital Services Coordinators[41] which are, in turn, designated by the European Commission.[42] In this regard, there are doubts as to the harmonization of practices throughout the EU, considering that there does not seem to be any indicator that there will be uniform choice criteria for each Member-State's coordinator. In this regard, a lack of uniformity in designating out-of-court dispute settling bodies indicates some frailties in the DSA's harmonization aims, considering its purposes as a regulator, which adds to the legal uncertainty already present in terms of the internal dispute resolution criteria mentioned above.

[40] Art. 14 as a provision mandating the codification of the Terms and Conditions will have an important on tackling discretionary and abusive content moderation practices, protecting fundamental rights.

[41] Art. 18(2) DSA.

[42] Art. 38 DSA.

Still on the topic of out-of-court dispute settlements, the DSA makes it clear that dispute settlement bodies do not have the power to impose a binding settlement on the involved parties.[43] This disposition constitutes a noticeable change from initial proposals of the DSA,[44] in which dispute resolution bodies' decisions on this topic were to be considered binding. The change stemmed from concerns of bad actors using these mechanisms to arbitrate every content removal decision at a company's expense and due to the risk of legal fragmentation.[45] While the legislator's concerns are legitimate for the purposes of the mechanism's effectiveness, it is relevant to question whether the change does not risk turning the mechanism more performative. Indeed, the changes raise the question of how effective this redress mechanism can be, considering how platforms can simply disregard any decisions reached out of court whenever they prove too unfavorable for them. Moreover, it puts the very functionality of the entire redress system into question, in the sense that platforms might deliberately opt to exhaust all available mechanisms until only judicial redress is left. Though users are theoretically able to resort to courts, in practice, the costs associated with legal proceedings can prove to be a barrier for content creators to go through with challenging removals or restrictions that they consider unfair.

There is an inherent power imbalance between hosting service providers and users, particularly in what concerns to resources. This imbalance only becomes more prevalent in the case of VLOPs which, despite the numerous obligations they are subject to by the DSA, seem to still have broad leeway in terms of their content control, regardless of how broad or restrictive their terms and conditions might become. Furthermore, these vulnerabilities raise concerns on the effectiveness of the DSA's efforts to combat whitelisting practices, which favor high-profile accounts with large numbers of followers when it comes to the protection of their speech (Gorwa et al. 2020). Indeed, considering what has been exposed, it is not far-fetched to consider that these accounts and their respective users will be the only ones monetized enough to be able to exercise the mechanism of judicial redress, not to mention that they might continue to benefit from greater leeway.

Overall, despite the changes made to ensure that dispute resolution bodies' decisions are not to be binding, the very validity of the inclusion of out-of-court dispute settlement processes in the DSA raises severe concerns. Scholars Wimmers 2021 have pointed out how this procedure is both unnecessary and counter-productive for the goals of the DSA and how its very theoretical conception undermines the value of normative decision usually reserved for the judiciary (Wimmers 2021).

---

[43] Art. 21(2) DSA.

[44] European Parliament (2022) Proposal for a Digital Services Act. https://www.europarl.europa.eu/doceo/document/TA-9-2022-0269_EN.html.

[45] This choice also is in line with the European Commission's initial goal of avoiding legal fragmentation through the DSA. See the Explanatory Memorandum (n 1) in European Commission (2020) 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.

5.1.5 Collective Actions

The last resort for solving disputes surrounding illegal content in these platforms is, naturally, judicial. The DSA does not attempt to subvert or change this fundamental principle. It aims to implement many instruments and mechanisms that should be used first (by third parties, users and even administrative authorities), from internal complaint systems to out-of-court dispute resolution. Nevertheless, as shown in the previous paragraphs, many of their provisions have several cracks through which problems may arise. Art. 54 provides that users can always request compensation for damages and damages caused by OSP that have breach obligations under the DSA.[46]

Unfortunately, the DSA does not include measures that aim to facilitate access to court for individuals or weaker parties when all else fails. Access to justice in these matters will still be regulated by national law, in compliance with the Charter of Fundamental Rights, namely art. 47.[47] It could be argued that these instruments were largely not designed for this kind of problems and in practice may not be sufficient to ensure the principles of effectiveness and equivalence.

Nevertheless, the DSA does include two provisions that aim to strengthen the position of individuals by promoting collective action, in different models: Arts. 86 and 90. Both provisions have the objective of allowing organisations, associations and other entities to exercise the rights of recipients of intermediary services conferred by the DSA in their representation. By pulling resources together and similar complaints of non-compliance and infringement, these entities may be able to pursue the interests of larger amounts of recipients that otherwise could not do it.

Art. 86 is broader in its subjective scope, allowing for any "type" of recipient of intermediary services to be part of these bodies. They may include natural persons such as consumers and other individual users (that do not qualify as such) and legal entities, from non-profit organisations, universities, businesses, act. However, the entities that are mandate for representation in accordance with Art. 86 cannot themselves be profit motivated and must pursue a legitimate interest that the DSA is complied with. This restriction is welcomed to ensure that this representation is not subverted, as it happened in past (Alexander 2019; Geyser 2022).[48] Recital 149 refers that the primary purpose of these entities will be the exercise of rights related with the submission of notices, challenging decisions by intermediaries and lodging complaints, while Art. 86(2) reinforces this, by stating that complaints submitted by these entities will be processed with priority and without undue delay.

Art. 90 represents a true final effort regarding the matter of access to justice in the DSA. It is an amendment that adds the DSA to the annex of Directive (EU) 2020/1828

[46] See Fn68: 13.

[47] See Fn68: 12.

[48] As it happened in the past with Multi-Channel Networks (MCN) on Youtube. MCN were organisations that were created to represent many individual youtubers when dealing with Youtube, but many became infamous due to predatory practices (such as revenue theft and perpetual contracts) against the people that they were supposed to represent. Additionally, many were sold to large media conglomerates.

on representative actions for the protection of consumers.[49] This simple provision allows private qualified entities to pursue injunctions and reparation for damages against OSP for violations of the DSA in behalf of large groups of consumers.

While it has a lot of distinctions from the American class-action model and it does not intend to promote the emergence of a litigation industry—which prioritizes the litigator's financial gains over consumer interests (Agulló 2022), this Directive aims to strengthen collective action with the clear objective of tackling concerns regarding the principle of effectiveness, access to justice and enforcement of European (consumer) law. It requires all Member States to incorporate representative actions for redress and injunctions in their legal systems, for national and cross-borders cases, but it allows for much discretion in the transposition, including in more controversial matters, such as the choice of opt-in or opt-out system.

In summary, despite Art. 90 being a welcomed addition to the DSA, its effectiveness is yet to be assessed. While in certain Member States with a more active civil society, it may provoke useful externalities by indirectly promoting compliance with the DSA; in in others it may have little to no effect. It is also necessary to refer that Directive 2018/1828 is limited to protect consumers, which leaves out of its subjective scope many individuals and entities which still require protection.

5.2 On the Topic of Transparency under the DSA

The very effectiveness of the mechanisms mentioned prior depends on a proactive approach from users. Indeed, creating specific bodies and rules to combat the "over-blocking" phenomenon is unlikely to effectively combat unjustified and excessive content restriction practices by itself if users do not actively resort to the proper mechanisms. This is only made possible if the users are cognizant of how their content is being affected by the moderation tools, which demonstrates the key role transparency plays in the DSA's goals of empowering users and in allowing for a proactive and meaningful (Mylly 2023) exercise of rights. It is therefore relevant to assess the role the DSA's transparency obligations play in empowering users and combatting "over-blocking", if at all.

In theory, the DSA obliges providers of intermediary services to fulfill minimum transparency and fairness requirements[50] in their T&C, in the sense that they are required to include information on any policies, measures, procedures and tools used for content moderation. Alongside the statement of the reasons for the content's removal—which include the legal or contractual ground behind the decision -, hosting service providers are to also include information on the available redress mechanisms. Depending on the specific case, affected recipients of the server may have access

[49] Directive (EU) 2020/1828 of the European Parliament and of the Council of 25 November 2020 on representative actions for the protection of the collective interests of consumers and repealing Directive 2009/22/EC (Text with EEA relevance).

[50] Examples of this are Recitals 50 to 54 and Arts. 15 and 16 DSA.

to internal complaint-handling mechanisms,[51] an out-of-court dispute settlement,[52] and/or judicial redress.[53]

All these measures towards transparency are in line with the DSA's obligations for all intermediaries to publish a yearly report detailing their content moderation operations,[54] further demonstrating an effort to distribute power through the internet more evenly (Genç-Gelgeç 2022). Moreover, online platforms must include information on their internal complaint handling systems in the form of reports, in which they must identify the number of disputes referred to out-of-court dispute settlement bodies, alongside the average time needed to complete said proceedings.[55] Indeed, whether these reporting obligations end up originating reports with strategically structured information remains to be seen, though there is an understandable risk that service providers can feel inclined towards embellishing their results to avoid scrutiny. The effectiveness of the transparency obligations for providers of intermediary services and of the impact of the annual transparency obligations regarding the report of their content moderation actions is still to be assessed in practice. For this purpose, the first year of the DSA in effect being essential to later enrich the following assessment.

The fact that the DSA provides for this type of specific transparency obligation on behalf of hosting service providers demonstrates the legislator's commitment to empower users regarding content moderation that might affect them. Given that the exercising of rights depends on people being aware that they are being affected by another agent, the legislator aids users in challenging decisions that might affect them in a timelier and more effective manner. In theory, by having affected recipients be told of the specific reasons for the removal or restriction of the content and on the available means of redress, the DSA allows them to specifically challenge the removal's reasoning in the route specifically projected for said purpose. It also constitutes an attempt at minimizing "shadow bans" to a certain degree, though the extent of this minimization with be very dependent on the practical application of the Regulation.

In practice, however, the obligations mentioned above do not necessarily ensure that users will be clarified on the criteria that providers resort to when blocking or removing content. Indeed, the DSA requires providers to inform users regarding the existence of their means for content moderation, but not on how they operate. Similarly, though providers must inform users on the reasons for the content's removal, they are not obliged to indicate how they reached their decision, how the reasons presented influenced the decision, nor on the amount of automation involved in said

[51] Art. 20 DSA.

[52] Art. 21 DSA.

[53] Though the DSA does not provide for this route in its articles, it recalls that it must be made available. Moreover, judicial redress is subject to specific national legislation and procedures, and as such, the affected recipients should resort to said national norms to challenge any decisions that might impact them.

[54] Art. 13 DSA.

[55] Art 23(1) DSA.

decision.[56] On that regard, the level of transparency of provider's T&C can be brought into question, and in turn, raise concerns on how well users can oppose decisions and the fairness (or lack thereof) of the criteria used to remove or restrict content.

Despite these concerns, the DSA does demonstrate a commitment to avoid performative transparency by always requiring disclosure of crucial information, which did not always happen under the ECD. If anything, even under the risk of subversive reporting, there is an inherent increment of the information that must be disclosed to users. Though these measures' practical implications have yet to be verified, this exercise, in theory, promotes broader knowledge of the decisions made about user content and, consequently, demonstrates an effort in empowering users.

5.3 On the Role of Digital Services Coordinators: An Analogous Comparison with Data Protection Authorities

Alongside the concerns presented above, it can be said that DSA proposal's formulations suggest that the regulation perceives content blocking as an interim measure, rather than a last resort. This is made clear in how the Digital Services Coordinators[57] are endowed with the capacity to order the blocking of websites, which can have a global impact for all users.

The lack of specific dispositions in what constitutes a violation risks making Digital Services Coordinators engage in the "over-blocking" phenomenon, rather than adopting a truly case-by-case approach in mitigating it. Indeed, for all intents and purposes, the Digital Services Coordinators' priorities, strategies, and stances are yet to be verified in practice. Moreover, one must not ignore that the various Member-States each hold varying legal traditions and doctrines that necessarily impact the way each Coordinator will perform its monitoring duties. This degree of uncertainty may prove disconcerting for OSP, who face fines of up to six percent of their annual global turnover upon eventual failures to comply with the DSA, and who may be over restrictive in their T&C to avoid these hefty fines.

Though it is difficult to predict how Digital Services Coordinators might exercise their monitoring powers, it can be possible to do so. Furthermore, in the context of digital platforms, it is relevant to look at the context of personal data protection supervisory authorities (DPAs) and on their specific fines towards big tech companies and platforms that constitute intermediaries and Very Large Online Platforms (VLOPs).[58] This analogy is relevant considering the relation between the DSA and

[56] Art. 15(1)(b)(e) DSA.

[57] Art. 38 DSA.

[58] On this, see Chapter IV, Sect. 4 DSA. VLOPs are a relevant focus of the DSA and are defined in Art. 33(1) of the DSA as being "online platforms which have a number of average monthly active recipients of the service in the Union equal to or higher than 45 million".

the General Data Protection Regulation (GDPR),[59] given that the DSA is meant[60] to be articulated with and complement the GDPR's dispositions.[61] Therefore, it can be interesting to assess how DPAs exercise their competences regarding non-conformity with the GDPR, in order to predict what might be expected from Digital Services Coordinators, given the DSA and the GDPR's close relation. As it stands, the highest GDPR fines applied to date mostly pertain to the practices of digital platforms, with companies and groups like Amazon,[62] WhatsApp[63] and Meta[64] being some of the most frequently sanctioned actors. Considering that all of these sanctioned actors are also intermediaries and VLOPs as per the DSA's definitions, it is not unfair to muse that the same level of scrutiny and high fines is to be expected from the recently appointed Digital Coordinators, who are likely to align with DPA's stances and practices towards certain actors.

With this context in mind, and as stated previously, intermediaries and other service providers (whether they are VLOPs or not) are unlikely to be willing to risk more hefty fines and reputation damage, based on which a more radical approach of defining very strict T&C simply removing or restricting content without many criteria is more likely. From a practical standpoint, it is technically simpler and safer for platforms to just wait for users to trigger redress mechanisms and only then handle the take-down's particularities. This may hinder the accuracy of the notice and action mechanisms and the creation of balanced moderation criteria. Indeed, the vast number of resources required to assess allegations under a notice-and action makes it unlikely that intermediaries other than VLOPs will be able to properly devote time and effort to effectively assessing notifications. These reservations also extend to internal redress mechanisms. It is, therefore, more likely for intermediaries to simply remove or restrict all notified content to save resources and avoid scrutiny rather than risk a lackluster attempt at rigorous moderation in the face of reduced resources.

Should the prior considerations turn out to be true, users might be disempowered due to being discouraged from even resorting to the redress mechanism both due to overly strict T&C and to no expectations of success in overturning decisions over

[59] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[60] European Commission (2022) Questions and Answers: Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348.

[61] Recitals 10, 34, 67, 68, 69, 71, 94 and 98 of the DSA clarify some of these complementarities.

[62] Amazon Europe was fined € 726 M in Luxembourg, though the execution of this fine has since been partially suspended by order of the president of the Administrative Court of Luxembourg. See Tribunal administratif du Grand-Duché de Luxembourg (2021) Audience publique du 17 décembre 2021 < https://justice.public.lu/content/dam/justice/fr/actualites/2021/46630ord.pdf.

[63] Irish Data Protection Commission (2021) Data Protection Commission announces decision in WhatsApp inquiry. https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-announces-decision-whatsapp-inquiry.

[64] Bryant, Jennifer (2023), Irish DPC fines Meta 390 M euros over legal basis for personalized ads, available at https://iapp.org/news/a/irish-dpc-fines-meta-390m-euros-over-legal-basis-for-personalized-ads.

their content. Naturally, the validity of these considerations greatly depends on the practical application of the DSA and of the Digital Services Coordinators' practices in their first years of existence, but as it stands, the current transparency obligations for content moderation do not appear to bode well for a fair content moderation framework. On this note, it will be easier to perform a more thorough assessment of the practical implication of Digital Services Coordinator's practices through the analysis of their respective annual reports, which are mandatory, as stated by the DSA.[65]

6 Conclusion

An effective liability framework needs to acknowledge the interdependence between all parties in this matter. It needs to carefully consider the underlying economic and social trade-offs. There are three essential elements.

1. The liability framework needs to encourage the participation of users, injured parties and OSP. It is structural requirement, aimed at reduction of the asymmetry of information, and coercing each to "do their part". OSP should have a duty of care to monitor and subsequently reduce expected harm. This can be further promoted through a "Good Samaritan clause". Rating, notice-and-takedown and flagging systems are essential, and third parties and users should be motivated to use them.
2. A liability rule should incentivize the diligent monitorization and removal of illegal content by OSP, with minimal errors. Strick one-sided liability will inevitably coerce OSP to engage in over-blocking and censorship. Rightsholders need to be held accountable for abuse and error, while users need a certain level of protection to mitigate chilling effects. If OSP detract themselves from the dispute process, leaving rightsholders to preside over it, the results will seriously harm fundamental rights and social welfare (Buiten et al. 2019).
3. Therefore, this liability regime should protect the OSP only when certain procedural obligations are implemented that ensure the transparency and accessibility of these systems. OSP should be required to be regularly audited and produce detailed reports to the public, about all the relevant statistics, from takedown-notices to complaints about over-blocking (such as the model in the German Network Enforcement Act). Moreover, it is "no one-size-fits-all" liability rule for all types of harm and OSP. Attempting it would likely amplify current asymmetries, creating serious market barriers which would ultimately promote the large incumbents, and would generate substantial inefficiencies (Lefouli et al. 2021).

[65] Art. 44 DSA.

The ECD's model does not need to be completely reinvented, but rather reformed towards an adaptable principle-based framework, that efficiently shares the responsibility of detection and removal among the many actors, with adequate checks-and-balances and enhanced transparency (Buiten et al. 2019). The DSA attempts to codify some of these ideas, though its success in practice remains to be assessed.

Efforts to mitigate the pernicious effects of disinformation[66] and hate speech tread a fine line with the protection of users' fundamental rights,[67] and any imbalance[68] in this exercise risks enlarging the risk of power asymmetries between users and the platforms that monitor their content. Legal initiatives should combat the growing lack of incentives users have to contest decisions made over their content. On this note, it remains to be seen whether the DSA's dispositions provide an improvement in the current state of the art. However, the DSA's dispositions prove to be both broad and ambiguous, which, when paired with the impacts of an out-of-court settlement in the redress procedure for content moderation, raise serious doubts on the effectiveness of this legislation in combating the asymmetries mentioned above and the chilling effect of the "over-blocking" phenomenon.

The DSA's goal should align with creating a framework that promotes the exercise of fundamental rights and free, democratic discourse, for which there are reservations as to how the legislator's choices will actually fulfill these purposes (Reda and Selinger 2021). Mitigating these concerns is key to fulfill the goal of making the DSA a global standard-setter[69] in terms of consumer protection, and as such, time will tell what the practical repercussions of the DSA will truly look like.

References

Agulló D (2022) The interplay of data, consumer and Private International law rules in the area of collective access to justice in the European Union, available at https://repositorio.comillas.edu/jspui/bitstream/11531/74625/1/Lainteraccionentrelasnormasdeproteccion.pdf

Alexander J (2019) These YouTubers are owed $1.7 million, and they're probably never going to get it/Ally Bank responds to concerns from creators. The Verge. https://www.theverge.com/2019/1/29/18202131/matpat-defy-media-youtube-ally-bank-ryland-adams-multi-channel-network. Accessed 29 Aug 2025

[66] These efforts are in line with the European Commission's European democracy Action Plan. See European Commission (2020) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions On the European democracy action plan COM(2020) 790 final. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0790&from=EN.

[67] On this topic, the Advocate General's statements regarding freedom of expression in Case C-401/19 are essential to consider.

[68] On this topic, see Joined Cases C-682/18 and C-683/18 Frank Peterson v Google LLC, YouTube LLC, YouTube Inc., Google Germany GmbH (C-682/18) and Elsevier Inc. v Cyando (C-683/18) (Opinion of Advocate General): 151.

[69] European Parliament (2020) Digital Services Act: Improving the functioning of the Single Market. https://www.europarl.europa.eu/doceo/document/TA-9-2020-0272_EN.html.

Appelman N (2023) Research Report on Disparate Content Moderation—DSA Observatory. DSA Observatory. https://dsa-observatory.eu/2023/10/31/research-report-on-disparate-content-moderation/

Bar-Ziv S, Elkin-Koren N (2018) Behind the scenes of online copyright enforcement: Empirical evidence on notice & takedown. Conn Law Rev 50:339

Buiten M, de Streel A, Peitz M (2019) Rethinking liability rules for online hosting platforms. Discussion Paper Series – CRC TR 224. 074:10–11. https://www.wiwi.uniunibonn.de/bgsepapers/boncrc/CRCTR224_2019_074.pdf

Calvi A, Kotzinos D (2023) Enhancing AI fairness through impact assessment in the European Union: a legal and computer science perspective. https://doi.org/10.1145/3593013.3594076

Desjardins J (2019, March 19) How the Tech Giants Make Their Billions. https://www.visualcapitalist.com/how-tech-giants-make-billions/

Dinwoodie GB (2017) A Comparative Analysis of the Secondary Liability of Online Service Providers. Secondary Liability of Internet Service Providers 1–72. https://doi.org/10.1007/978-3-319-55030-5_1

Erickson K, Kretschmer M (2018) What motivates takedown of usergenerated content by copyright owners? evidence from the removal of music video parodies on youtube. J Intellect Prop Inf Technol E-Commer Law 9(1):75–89

European Commission (2014) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions towards a thriving data-driven economy. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014DC0442&from=EN

European Commission (2017) Working document on the free flow of data and emerging issues of the European data economy COM (2017) 9 final. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52017SC0002&from=EN

European Commission (2020) European Commission 2020 work programme: An ambitious roadmap for a Union that strives for more. https://ec.europa.eu/commission/presscorner/detail/en/ip_20_124. Accessed 29 Aug 2025

European Parliament (2022) European Parliament legislative resolution of 5 July 2022 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 (COM(2020)0825 – C9- 0418/2020 – 2020/0361(COD)). https://www.europarl.europa.eu/doceo/document/TA-9-2022-0269_EN.html

G'sell, F (2023) The Digital Services Act (DSA): A General Assessment. Content Regulation in the European Union – the Digital Services Act, Trier Studies on Digital Law, Volume 1, Verein für Recht und Digitalisierung e.V., Institute for Digital Law (IRDT), Trier April 2023, Available at SSRN: https://ssrn.com/abstract=4403433 or https://doi.org/10.2139/ssrn.4403433

Genç-Gelgeç B (2022) Regulating Digital Platforms: Will the DSA Correct Its Predecessor's Deficiencies?, 18 CYELP 25, available at https://www.cyelp.com/index.php/cyelp/article/view/485

Geyser W (2022) What Are MCNs for YouTube Creators (+ Top Multi-Channel Networks). https://influencermarketinghub.com/mcn-youtube-creators. Accessed 29 Aug 2025

Gillespie T. (2017) Regulation of and by Platforms, in "The SAGE Handbook of Social Media", available at https://sk.sagepub.com/hnbk/edvol/the-sage-handbook-of-social-media/chpt/14-regulation-and-by-platforms

Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7(1). https://doi.org/10.1177/2053951719897945

Haimson OL, Delmonaco D, Nie P, Wegner A (2021) Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. Proc ACM Hum Comput Interact 5(CSCW2):1–35. https://doi.org/10.1145/3479610

Hill RK (2016) What an Algorithm Is. Philos Technol 29(1):35–59. https://doi.org/10.1007/s13347-014-0184-5
Hornik J, llera V (2017) An Economic Analysis of Liability of Hosting Services: Uncertainty and Incentives Online. Bruges European Economic Research Papers 37/2017—Archive of European Integration. Pitt.edu. http://aei.pitt.edu/92487/1/beer37.pdf
Husovec M (2017) Injunctions Against Intermediaries in the European Union: Accountable But Not Liable?. Cambridge University Press. Sartor, Giovanni (2017) Providers Liability: From the eCommerce Directive to the future. Study for the European Parliament
Kaiser K, Ratcliff C (2022) Digital Services Act and Digital Markets act: Opportunities and Challenges for the Digital Single Market and Consumer Protection
Kaye DBV, Gray JE (2021) Copyright Gossip: Exploring Copyright Opinions, Theories, and Strategies on YouTube. Social Media + Society, 7(3), 205630512110369. sagepub. https://doi.org/10.1177/20563051211036940
Kuczerawy A (2022) Remedying Overremoval: The Three-Tiered Approach of the DSA. Verfassungsblog. https://doi.org/10.17176/20221103-215534-0
Morais Carvalho, J, Arga e Lima F, Farinha M (2021, May 24) Introduction to the Digital Services Act, Content Moderation and Consumer Protection. Papers.ssrn.com. https://ssrn.com/abstract=3852280
Mylly UM (2023) Transparent AI? Navigating Between Rules on Trade Secrets and Access to Information. Int Rev Intellect Prop Compet Law. https://doi.org/10.1007/s40319-023-01328-5
Nordemann JB (2018) Liability of Online Service Providers for Copyrighted Content – Regulatory Action Needed? In-Depth Analysis for the IMCO Committee of the European Parliament, European Parliament
Penney JW (2019) Privacy and Legal Automation: The DMCA as a Case Study, pp. 412–486, available at https://law.stanford.edu/publications/privacy-and-legal-automation/
Quintais JP (2020) The New Copyright in the Digital Single Market Directive: A Critical Look. Eur Intellect Prop Rev 42(1):28–41. https://doi.org/10.2139/ssrn.3424770
Quintais JP, Mezei P, Harkai I, Vieira Magalhães J, Katzenbach C, Schwemer SF, Riis T (2022) Copyright Content Moderation in the EU: An Interdisciplinary Mapping Analysis. SSRN Electron J. https://doi.org/10.2139/ssrn.4210278
Quintais JP, Schwemer SF (2022) The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright? Eur J Risk Regul 1–31. https://doi.org/10.1017/err.2022.1
Reda F, Selinger J (2021) Digital Services Act: European Parliament discusses website blocking against platforms. Verfassungsblog. https://doi.org/10.17176/20211118-202001-0
Sartor G, Loreggia A (2020) Study: The impact of algorithms for online content filtering or moderation ("upload filters") (pp. 1–69). https://doi.org/10.2861/824506
Savin A (2022) Designing EU Digital Laws. Copenhagen Business School. CBS LAW Research Paper No. 22–13. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4293314
Savora D et al (2022) The Digital Services Act—What it is and what impact will it have?, available at https://www.cliffordchance.com/content/dam/cliffordchance/briefings/2022/12/the-digital-services-act-what-is-it-and-what-impact-will-it-have-updated-december-2022.pdf
Streel A et al (2020) Online Platforms' Moderation of Illegal Content Online Law, Practices and Options for Reform, available at https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf
Urban JM, Karaganis J, Schofield B (2017) Notice and takedown in everyday practice. UC Berkeley Public Law Research Paper No. 2755628
Wimmers J (2021) The Out-of-court dispute settlement mechanism in the Digital Services Act: A disservice to its own goals. J Intell Prop Info Tech Elec Com L 12:381
Zeybek B, Van Hoboken J, Buri I (2022) Redressing Infringements of Individuals' Rights Under the Digital Services Act—DSA Observatory. https://dsa-observatory.eu/2022/05/04/redressing-infringements-of-individuals-rights-under-the-digital-services-act

Zingales N (2023) The DSA as a Paradigm Shift for Online Intermediaries' Due Diligence Hail To Meta-Regulation. In Putting the Digital Services Act into Practice Enforcement, Access to Justice, and Global Implications (pp. 216–218). Verfassungsbooks