Institutiones Administrationis – Journal of Administrative Sciences
DOI: 10.54201/iajas.v3i2.85

The Digital Services Act and the Problem of Preventive Blocking of (Clearly) Illegal Content

Marcin Rojszczak*

* Assistant Professor, Faculty of Administration and Social Sciences, Warsaw University of Technology, Warsaw, Poland. E-mail: marcin.rojszczak@pw.edu.pl

Abstract

The adoption of the long-awaited Digital Services Act (DSA) is undoubtedly one of the more significant successes related to the implementation of the ambitious EU Digital Strategy. In addition to important announcements that the new law will help to transform the next few years into Europe's digital decade, an update of the liability framework for digital service providers also provided an opportunity for a broader reflection on the principles of building governance in cyberspace. Indeed, the notice and takedown model, which had been in place for more than two decades, had become progressively eroded, leading service providers to increasingly implement proactive content filtering mechanisms in an effort to reduce their business risk. The aim of this article is to explore those changes introduced by the DSA which affect the regulatory environment for the preventive blocking of unlawful online content. In this respect, relevant conclusions of the ECtHR and CJEU jurisprudence will also be presented, as well as reflections on the possibility and need for a more coherent EU strategy with respect to online content filtering. The analysis presented will focus on filtering mechanisms concerning mainly with what is referred to as clearly illegal content, as the fight against the dissemination of this type of speech, often qualified under the general heading of "hate speech", is one of the priority tasks for public authorities with respect to building trust in digital services in the EU.

Keywords

Digital Services Act, illegal content, liability of intermediate service providers, content blocking

1 Introduction

"The freedom of expression exercised on online forums by anonymous authors often provokes unbridled speech that degenerates into hate speech violating the personal rights of third parties. The availability of hateful comments online can be virtually indefinite, and holding individual internet users accountable is in practice impossible."[1]

The statement presented above, formulated by the Polish Supreme Court, aptly summarises the complicated legal position of individuals in their efforts to protect their good name

[1] Polish Supreme Court, 30 September 2016, I CSK 598/15, www.sn.pl.

against untrue and often defamatory statements published online. The lack of effective tools for responding to such violations, combined with the complexity of the judicial route and the cross-border nature of the disputes, leads to a widespread belief that the existing legal model – including the framework governing the operation of service providers – does not protects the legal interests of individuals harmed by unlawful statements made online.

The problem of the use of online services – in particular hosting services – to distribute unlawful (illegal) content has been the focus of the EU legislature's attention for over two decades. One specific area concerns the determination of the liability of the so-called 'intermediary service providers' for actions carried out by users, in particular actions relating to the transmission of illegal material. This issue was first regulated in the e-Commerce Directive,[2] adopted in 2000, which established a general model for the liability of digital service providers. In particular, it introduced the so-called notice and takedown model, according to which the liability of a service provider is triggered once it fails to take action against specific content upon becoming aware of its unlawful nature (Julià-Barceló & Koelman, 2000).

For almost two decades, the notice and takedown model set the limit of responsibility of digital service providers for the actions of users online, including for the content they published. Significantly, the e-Commerce Directive simultaneously introduced the so-called prohibition of a general monitoring obligation – and thus opposed the adoption of national legislation requiring service providers to monitor all transmitted information for infringing content. This principle was interpreted over the years as a barrier to establishing a legal obligation to use automatic content filtering mechanisms.[3]

However, the massive growth of digital services has made the dissemination of illegal content a widespread phenomenon, the scale of which has become a real problem affecting users' trust in online services and is also slowing down the digital transformation process. As a result, in series of judgments the European courts – both the European Court of Human Rights (hereinafter: ECtHR) and the Court of Justice of the European Union (hereinafter: CJEU) – have clarified the conditions for recognising the liability of service providers, favouring a broader interpretation of content filtering obligations (Spindler, 2017). In the Delfi judgment of 2015, the ECtHR held, that a provider of a hosting service that carries out its activities on a professional and profit-making basis and uses technical means of content moderation to do so cannot be exempted from liability simply because the victim did not inform it of the infringement taking place.[4] This interpretation was provided against the background of a case in which the unlawfulness of the challenged postings was obvious, and therefore did not require legal analysis or detailed substantive knowledge. In this way, the ECtHR confirmed that statements that affront human dignity or incite physical violence not only do not deserve legal protection, but also require an adequate response from the judiciary, as well as service providers (Cox, 2014).

Indeed, making the assumption that an aggrieved person must use a particular service – or online services in general – when determining the extent of a service provider's liability would lead to a situation where, in the absence of notification, the service provider could not be held

[2] 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, OJ 2000 L 178/1.

[3] Nowadays, the prohibition of a general monitoring obligation is also increasingly referred to "as a general principle of law governing the Internet" – *see e.g.* Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Case C-401/19, EU:C:2021:613, para. 116; see also Wilman (2022).

[4] *Delfi AS v. Estonia* App no 64569/09 (ECtHR, 16 June 2015), para. 159.

liable for the dissemination of hateful or flagrantly hurtful speech. Based on this observation, the ECtHR held that in the case of manifestly illegal content, a service provider is obliged to act promptly – irrespective of whether or not it has received an additional notification on the matter. The CJEU accepted this view in its subsequent case law, while at the same time favouring the possibility of applying "extended orders" for the removal of content – obliging the provider not only to block the posts considered unlawful, but also to take the necessary technical measures so that "identical" or "equivalent" content cannot be published in the future (Rauchegger & Kuczerawy, 2020).[5]

Although the interpretation in the Delfi judgment has been further developed in subsequent ECtHR cases,[6] the general principle that a service provider has a duty to take steps against manifestly illegal content has become one of the cornerstones of the European model for regulating the liability of intermediate service providers (Spano, 2017). However, upholding this liability regime requires the use of increasingly sophisticated content moderation and filtering mechanisms, which on the one hand are supposed to identify instances of unlawful speech, and on the other hand must not lead to undue interference with freedom of expression – ultimately becoming a mechanism of preventive censorship (Bloch-Wehba, 2020).

The Digital Services Act (hereinafter: DSA or 'Regulation'), adopted in 2022, was intended to carry out a profound reform of the notice and takedown model that had been in place for 20 years. The intention of the European Union legislature was to strengthen the rights of users by, among other things, indicating more clearly the limits of service providers' responsibility. Therefore, the DSA expands the previous responsibilities and establishes new requirements in the area of online content filtering. The EU legislature's intention was to shift some of the responsibility for ensuring the legality of published content onto service providers, without losing sight of the need to respect freedom of expression and the right to information. After all, not every published content is illegal, and establishing an overly oppressive model could have a chilling effect on service providers, who might exercise overly extensive control of publications in order to reduce the risk to their business (Wu, 2013).

The purpose of this article is to discuss the most important changes proposed in the DSA which shape the rules for service provider liability in the area of automatic content moderation. In particular, it will explore the provisions concerning the application of preventive measures, i.e. those used for the preliminary screening (that is prior to publication) of content uploaded by users. For some years, the application of such measures has raised justified doubts, both about the effectiveness of the technical solutions being implemented, and the impact of such solutions on users' rights.

2 Broad definition of illegal content

In principle, both the notice and takedown model established under the e-Commerce Directive and the notice and action model introduced in the DSA link service provider liability to the service provider's knowledge of the unlawful nature of the information being disseminated

[5] Judgment of 3 October 2019, *Glawischnig-Piesczek v Facebook,* C-18/18, EU:C:2019:821, para. 39.

[6] See e.g., *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary* App no 22947/13 (ECtHR, 2 February 2016); Pihl v. Sweden App no 74742/14 (ECtHR, 9 March 2017); *Jezior v. Poland* App no 31955/11 (ECtHR, 4 June 2020).

(Kuczerawy, 2020). The service provider has an obligation to remove or to block access to content where the unlawfulness of the content has been established but also where it was merely plausible. Only if the service provider remains passive while knowing the unlawful nature of the content and does not remove the questionable publications can its (civil) liability for the consequences of the distribution of the illegal content be enforced. The general mechanism introduced by the e-Commerce Directive (and now by the DSA) thus serves to limit the liability of online intermediaries (in particular, hosting service providers[7]) for content stored by end users. Importantly, the notice and takedown model is not unique to the EU legal system but is also widely used in other jurisdictions throughout the world (Wang, 2018).

Knowledge regarding the unlawful nature of content can come from a notification received from a user, be the result of the work of the service provider's own employees (e.g. content moderators) or originate from automatic filtering systems (the so-called upload filters). In any case, the removal (blocking) of content should be limited to materials whose unlawful nature is obvious or has been priorly demonstrated.

As a result, an issue of major practical importance for the application of the obligations under the DSA is a clear definition of the concept of illegal content. Under the e-Commerce Directive – and the national laws implementing it – this concept had been interpreted in the context of the legal model of the individual Member States. This meant that the same content could be considered lawful in some jurisdictions and illegal in others. This was particularly true for extremely critical content, approaching the limits of defamatory or insulting speech. The differences in the legal categorisation of this type of speech were perfectly illustrated by the discussion sparked by the recent ECtHR judgment in *Sanchez v France,*[8] in which the Court held that the application of criminal sanctions against a Facebook user for content published in his news feed by other users does not violate the guarantees under the Convention, including in particular the right to information (Lemmens, 2022).

Indeed, it seems not enough to harmonise the procedural part of content filtering rules if their application depends on an external definition that has not been harmonised and can be determined separately in each country. In this case, the problem may not even be the existence of multiple definitions of illegal content but the lack of consistency between them. What may qualify as illegal content in one Member State may in another be a legitimate statement, the removal of which will lead to infringement.

With the entry into force of the DSA, its mechanisms which set the limits of intermediate service providers' liability are to be applied directly. As a result, the lack of a common definition of illegal content may lead to a reduction in the effectiveness of the new regulation and, consequently, it may also negatively affect the protection of users' rights.

As a general rule, the legal definition of illegal content in Article 3(h) of the DSA indicates information which, either in and of itself or in relation to a specific activity, is incompatible with EU law or the law of any Member State. However, under the Regulation the benchmark for

---

[7] One of the changes introduced in the DSA – compared to the e-Commerce Directive – is a greater nuance with regard to the responsibilities of individual intermediary service providers for users' actions. As with the e-Commerce Directive, the DSA, in principle, limits the application of the notice and takedown model to cases of the provision of hosting (data storage) services. At the same time, it also introduces additional obligations for hosting providers – related to explaining the reasons for the decisions made, including with regard to the moderation of content (more on this in further sections of this article).

[8] *Sanchez v. France* App no 45581/15 (ECtHR, 2 September 2021).

assessing the illegality of content can only be those rules of national law which are compatible with EU law. This therefore excludes the recognition as illegal content under the DSA of materials that have been declared as such on the basis of national rules incompatible with EU law.

According to the Regulation, illegal content is that which directly contravenes criminal law provisions laid down in EU acts – such as those indicated in Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law. The unlawfulness (illegality) of the publication may furthermore derive from the very nature of the information disseminated – e.g. relating to the abuse or exploitation of children (see the definitions in Directive 2011/93[9]); the promotion of extremist behaviour of a terrorist nature (see the definitions introduced in Directive 2017/541[10] and Regulation 2021/784[11]); as well as the infringement of copyright (see Directive 2019/790[12]).

In addition to the above, the unlawfulness of the content may also be a consequence of specific sectoral regulations, in particular regarding the unacceptability of certain practices used in online advertising[13] or the prohibition of advertising (promotion) of certain types of products.[14]

Importantly, to date no legal definition of hate speech has been developed in EU law.[15] Although the concept is commonly used in policy documents[16] as well as in drafts of legislation,[17] in reality the assessment of the inadmissibility (illegality) of certain publications due to their dissemination of so-called 'hate speech' must be carried out on the basis of existing regulations on the prohibition of dissemination of xenophobic speech and speech promoting racism (Brown, 2018). It can also be derived – albeit in an incomplete form – from the existing jurisprudence of the ECtHR and the CJEU, in which both Courts have decided on the inadmissibility of disseminating information that violates dignity or incites hatred or violence directed against a specific person or group of people. Hate speech is also examined in international law as a concept linked to the most serious crimes (Fino, 2020).

---

[9] Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA, OJ 2011 L 335/1.

[10] Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA, OJ 2017 L 88/6.

[11] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, OJ 2021 L 172/79.

[12] 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, OJ 2019 L 130/92.

[13] In this context, see the recent German Federal Office of Justice actions against Twitter's advertising practices, BfJ press release available at: https://cli.re/KJMZJa, accessed on 6 April 2023.

[14] See e.g., Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to medicinal products for human use, OJ 2001 L 311/67.

[15] However, as Teršek (2020) notes, a descriptive definition of "hate speech" can be found in the ECtHR case law.

[16] Communication from the Commission to the European Parliament and the Council – A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime, 9.12.2021 COM(2021) 777 final.

[17] See e.g., Draft proposal for a directive of the European Parliament and of the Council on combating violence against women and domestic violence, COM/2022/105 final.

In 2020, the European Commission announced the commencement of work on EU legislation harmonising national criminal laws against hate speech on the internet (Peršak, 2022). However, so far no draft of this type of regulation has been presented, which taking into account the European Parliament's election calendar and that the current Commission is nearing the end of its term of office, leads to the conclusion that this issue will not see a quick resolution at the level of EU law.

It should be stressed that in public debate, the term "hate speech" is most commonly used to characterise speech that goes beyond the mainstream discussion, often qualified as exclusionary or characteristic of extremist groups (Paz et al., 2020). However, the essence of introducing a legal definition of hate speech at the EU level is to single out that part of hate speech whose unlawful nature is beyond doubt. The aim of introducing this definition should, therefore, not be to cover all manifestations of hate speech, but only those which, by their very nature, are likely to be legally unacceptable in all Member States.[18]

In addition to the assessment of the unlawfulness of a publication made on the basis of binding EU law, Member States also have a great deal of latitude in shaping the relevant national rules – in particular with respect to defamatory or insulting content, but also in relation to particular categories of publications (e.g. offences against religious sensitivities). These rules may be enacted both in areas excluded from EU law (e.g. the disclosure of state secrets) as well as in areas within the scope of application of EU law where the Union has not exercised its competence.

Individual Member States – recognising the gaps in EU regulations against the dissemination of illegal content – have in recent years also started to adopt their own national legislation aimed at combating hate speech more effectively (Hochmann, 2022). An example of such a national regulation is the French law against online hate speech (also known as the Avia Law, named after its drafter).[19] In principle, the law only covered large information society service providers.[20]

The Avia Law introduced specific obligations to report and block online content and included its own definition of illegal content. The substantive scope of the act covered content violating human dignity, including hate speech or insults based on race, religion, ethnicity, sexual orientation or disability. Similar regulations have also been adopted in some other Member States, including Germany and Austria (Echikson & Knodt, 2018; Griffin, 2022).

[18] One example is hate speech against LGBT communities, which is not a publicly prosecutable offence in every Member State. A case in point is Poland: under Poland's criminal code, which has been the subject of criticism for a number of years, the prosecution of the offence of publicly insulting a person or a group of persons requires that the perpetrator's action be motivated by the "national, ethnic, racial or religious affiliation or lack of religious affiliation" of the victim(s) (see Art. 257 of the Polish Criminal Code). Similar restrictions do not exist in most of the other Member States, see Molter (2022, 6).

[19] Loi n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet (Act n° 2020-766 of June 24, 2020 aimed at combating hateful content on the internet); version of the act that was later subject to constitutional review available at: https://cli.re/1daYrb.

[20] Article I-2 of the Avia Law.

3 Preventive content blocking

In principle, the DSA establishes two mechanisms that may require content filtering measures to be applied in a preventive manner.

The first, stemming from Article 9 of the DSA, relates to the obligation to implement the so-called "orders to act against illegal content". Orders of this type, issued by Member States' authorities empowered to do so (not necessarily judicial ones), can oblige a service provider to remove the publications indicated therein and to implement measures against the re-publication of identical or equivalent content. The possibility of applying such a measure has already been confirmed by the CJEU in the Glawischnig-Piesczek case, the which concerned of which was the permissibility of ordering a service provider to actively search for illegal content conforming to a previously notified search-pattern (Cavaliere, 2019). The Court held that such an obligation does not violate the prohibition laid down in Article 15(1) of the e-Commerce Directive – that is, the prohibition against establishing a general monitoring obligation. In the Court's view, if the service provider is informed of specific instances of illegal content, the search for repetition of that content (or equivalent information) still falls within the logic of the liability framework under the notice and takedown model – and thus serves to eliminate published content whose illegal nature has become known to the provider (Keller, 2020). This is a highly controversial interpretation, which seems to ignore the reason for the prohibition against a general monitoring obligation under EU law. After all, the reasoning for that prohibition was based on recognising that a service provider should not be an actor actively involved in seeking out illegal content, and consequently should not become a general censor of users' activities, becoming the sole arbiter of the direction, scope and purpose of the monitoring of others' speech.

In addition to creating an obligation to implement measures to ensure the effectiveness of the orders received, the Regulation also establishes far-reaching obligations regarding the application of preventive content filtering mechanisms by the so-called 'very large online platform providers'. Basically, this group may include entities whose services are used by at least 45 million users within the European Union. The Regulation also establishes additional obligations for this group of entities, including assessment and minimisation not only of the systemic risk associated with the use of the service provided for the distribution of illegal content, but also of its negative impact on the protection of fundamental rights – such as respect for dignity; the right to privacy; protection of personal data; freedom of expression; non-discrimination; and the protection of the rights of minors. Service providers for whom a systemic risk assessment reveals a significant risk are required to implement "reasonable, proportionate and effective" mitigation measures. Among these, the Regulation mentions, inter alia, "adapting content moderation processes", which involves in particular, a reduction in response times and improving the quality of decision-making. Although the Regulation does not explicitly indicate an obligation to use algorithmic systems for this purpose, the reality of large service providers is that it is not possible to comply with this requirement using manual content screening. This applies, in particular, to the identification of harmful material, e.g. related to cyberbullying or child sexual abuse. Therefore, in some cases – like those related to instances of incitement to hatred – compliance with the Regulation may require the establishment of procedures and systems that make it possible to block the content considered unlawful within a maximum of 24 hours.[21]

[21] Cf. recital (87) of the DSA.

It is worth recalling that the DSA establishes a general model of liability, but the obligations introduced are subject to clarification in provisions adopted as lex specialis. Examples are Directive 2019/790 on copyright and related rights in the Digital Single Market (hereinafter: CDSM Directive), and Regulation 2021/784 on addressing the dissemination of terrorist content online (hereinafter: Terrorist Content Regulation). Significantly, in the case of the obligations indicated in the latter act, – service providers must take action within 60 minutes of receiving an order (Rojszczak, 2023).

Given that hundreds of millions of pieces of content are published in electronic media every day, it is clear that service providers cannot meet the high demands associated with countering unlawful publications without extensive use of automatic content filtering systems (Gillespie, 2020). According to statistics published by X (formerly: Twitter), in July-December 2021 the company removed over 5 million pieces of infringing content.[22] At the same time however, not in all areas of application can algorithmic systems identify cases of abuse with equal effectiveness. For example, according to Facebook in 2018 only 38% of removed material qualifying as hate speech was identified automatically, while for content promoting terrorism the figure was 99% (Koebler & Cox, 2018).

The above illustrates that one of the challenges faced by the EU legislature when drafting the DSA was to balance the application of automatic mechanisms in such a way that, while ensuring their effectiveness, they would not excessively interfere with users' rights. However, it is difficult in fact to prejudge whether this objective has been properly achieved, as the final version of the Regulation addresses this issue in a very general manner. The DSA imposes a number of detailed obligations on service providers, but without defining clear standards to assess whether the measures applied go beyond what is necessary. Automatic moderation of content may be applied both at the stage before publication (upload filters) and to content already published to which users have raised objections. It seems that in the former case – i.e. blocking of information at the pre-publication stage – the act should have at least provided for the use of fast-track complaint mechanisms in order to ensure that a user's complaint can be dealt with within a reasonable time and without the application of automatic measures.

In fact, the Regulation does not even regulate in detail the areas in which upload filters can be applied. Although in principle a service provider is obliged to apply measures that respect the rights of users and are proportionate to the pursuit of the intended and legitimate aim, the Regulation does not at the same time set a clear standard to assess how much content misclassification can be considered acceptable.

Although it may seem impractical to define such limitations directly in the Regulation, indicators of that kind could be included in lower-level legislation, for example in the form of certification schemes for the correct operation of content filtering algorithms.[23] In any case, establishing measures that interfere with fundamental rights without defining a clear framework for their application creates risks for both users and service providers.

This issue seems all the more problematic because according to Article 7 of the DSA, a service provider may not be denied the benefit of a limitation of its own liability in cases where, by its actions "in good faith and in a diligent manner" it takes measures which, in its judgement,

[22] See the "Rules Enforcement" section of Twitter's transparency reports. Online: https://tinyurl.com/4u8pp3sj

[23] Similar mechanisms have been implemented for several years as part of the work related to the enactment of Regulation 2019/881 (the Cybersecurity Act) and the establishment of an EU certification framework for technologies used in the area of cybersecurity (Fuster & Jasmontaite, 2020; Mitrakas, 2018).

serve to detect, identify, or remove the content. This provision de facto establishes a general clause, substantially limiting the possibility of holding service providers liable for instances of excessive content censorship.

Preventive content blocking may take place based on either the information indicated in the notification (order), or as a result of an analysis of the forthcoming publication in cases where its illegality is evident given the wording used and/or the content presented.

In the first case, the service provider relies in fact for its decision-making on information provided to it by external parties – e.g. copyright owners or trusted parties identifying certain categories of illegal material (e.g. child paedophilia or terrorist content). Contrary to appearances, in this case too the service provider's role will not be limited to a simple – and technically uncomplicated – search for identical publications (e.g. by comparing digital identifiers generated by one-way hash functions, because in this way it would only be possible to identify binary identical publications, but not "equivalent" ones, as expected by the CJEU.

Multimedia material, even slightly processed (e.g. in terms of colours, duration, addition of a watermark, etc.) will no longer be binary identical. In such a case, algorithms more complex than mere hash functions need to be used to confirm that the content under examination is "equivalent" to another publication previously deemed unlawful. An example of such an algorithm is PhotoDNA, developed and used by Microsoft for, inter alia, the identification of paedophile content (Lee et al., 2020). Similar systems are being developed to identify hate speech. In this case, standard dictionary-based systems are also gradually being replaced by state-of-the-art text classification algorithms (Siegel, 2020).

The majority of such systems are subject to copyright protection, and the details of their operation are secret. In principle, there is no doubt that such solutions are based on machine learning algorithms (Isaac et al., 2022). This means that their effectiveness largely depends on test data – and therefore on the quality of the decisions made beforehand. In such a case, the decision to block content is not based on the compatibility of the material examined with previous material whose illegality has already been established, but on an individual, automatic assessment of the publication in question, resulting in the assignment of a specific risk index. Depending on the decisions taken by the service provider, material above a certain risk threshold is qualified as infringing the terms of service, and its publication is stopped. In reality however, this does not mean that the material contains illegal content (or is itself unlawful). It means nothing more than (and nothing less than) that the algorithm has determined, with a high degree of certainty, that the analysed material is similar to publications previously assessed as unlawful. In the case of a specific, individual decision, such a conclusion is not necessarily true in every case – after all, statistical validity does not predetermine the accuracy of individual decisions (Llansó, 2020). Therefore, in practice, automated systems are used to first block material that is likely to contain illegal content. This material is then reviewed in a manual procedure (Gorwa et al., 2020, 6).

The way in which the rules of service providers' liability are defined in the DSA means that these entities do not actually bear the negative consequences of the use of overly elaborate content filtering mechanisms. Thus, while obvious cases of abuse – such as the intentional blocking of a particular publication – would give rise to a breach of the Regulation, the use of elaborate algorithmic mechanisms – which often operate on the basis of non-transparent algorithmic models and may ultimately lead to the blocking of the same content – would not have negative consequences for the service provider.

The discretionary nature of decisions taken by service providers is limited in the DSA by several mechanisms. First, for certain types of services (in particular, hosting services), a

detailed explanation of the reasons for the decision must be provided, including, among other things, the reasons for blocking access to a particular publication. The second is a tiered redress mechanism, which – depending on the type of service and the size of the service provider – may include internal complaint-handling mechanisms, out-of-court dispute settlement and judicial redress. In practice, however, the DSA does not set any requirements as to the length of time taken for the resolution of such complaints, with the result that while the blocking (removal) of content may be an action carried out very quickly, the reinstatement of the contested publication may even take many years. In such a case, the judicial protection of freedom of expression may not be sufficient to protect users from overly broad content-blocking measures.

Finally, the third type of measure relates to the transparency of the decisions taken, ensured by extensive reporting mechanisms. While on paper the DSA obliges service providers to report detailed statistics showing the effectiveness of their content moderation measures, including the number of publications blocked as well as the number of complaints filed and the number of successful complaints, in reality this information – which is supposed to ensure the transparency of the measures taken – leaves little room for comparison between the practices of different service providers (Leerssen, 2023; Parsons, 2019). This problem was clearly illustrated by the summaries submitted by service providers in accordance with Regulation 2021/1232[24] – the level of detail of which is far from sufficient to draw conclusions about the quality of the content analysis mechanisms implemented (in this case in relation to electronic communications).

4 Future regulatory strategies for preventive content filtering

The discussion on the new obligations, and consequently on the new powers for service providers reveals an increasing regulatory trend towards a gradual transfer of responsibility for cyberspace governance to private actors (Kikarea & Menashe, 2019). On the one hand, such an approach seems to solve existing problems, often related to the slowness of public authorities in identifying and responding to cases of publication of illegal content; but on the other hand, it creates a risk of excessive censorship and a real loss of control over the activities of big tech companies, which may use their own non-transparent algorithms. These risks were already apparent before the DSA came into force and have led to discussions on alternative forms of liability regulation for digital service providers, including outside the EU (Klonick, 2018).

Interestingly, inspiration in this respect can be found in other legislative initiatives by the European Commission. The proposed Regulation to Prevent and Combat Child Sexual Abuse (hereinafter: CSAM Regulation)[25] proposes a new – partly extended – coordination scheme for content removal processes, in which the tasks performed by service providers are carried out under the closer supervision of public authorities. In particular, the draft regulation establishes a new type of legal measure – the so-called ‘detection orders’[26] – to set rules to be applied

[24] Regulation (EU) 2021/1232 of the European Parliament and of the Council of 14 July 2021 on a temporary derogation from certain provisions of Directive 2002/58/EC as regards the use of technologies by providers of number-independent interpersonal communications services for the processing of personal and other data for the purpose of combating online child sexual abuse, OJ 2021 L 274/41.

[25] Proposal of 11 May 2022 for a regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse, COM/2022/209 final.

[26] See Art. 7 of the draft regulation (n. 26).

by the online service provider for the filtering (analysis) of content. While the provider will still be left free to decide on the technical measures to be applied, the order will define key criteria in order to limit the intrusiveness of the measures taken and ensure that they "do not go beyond what is strictly necessary to effectively address the significant risk identified."[27] However, above all, the establishment of detection orders could help resolve doubts about how to legally qualify the filtering mechanisms used by service providers. In the sense of the proposed regulation, these activities will be performed by private entities, but on behalf of and under the supervision of public authorities and for the performance of a public task. As a result, doubts about whether these practices should be assessed in a similar manner to other forms of surveillance implemented by public authorities will disappear. In turn, this will create scope for an examination of the legality of the actions taken in the context of the surveillance standards shaped in the rich jurisprudence of the ECtHR and/or the CJEU.

The draft CSAM Regulation is noteworthy for another reason. It will set up an EU coordination body (the EU Centre on Child Sexual Abuse), one of whose tasks will be to maintain a register of reports (database) containing information on detected paedophile content. This register could be used in the future as a source of reliable information on illegal material that should be blocked by service providers in individual Member States. This mechanism creates a promising way to manage the process of blocking online content. The centralisation of information concerning the actions taken by individual service providers simultaneously increases the efficiency of content blocking mechanisms and the transparency of the decisions taken. Moreover, the proposal provides for content filtering technologies developed by the Observatory to be made available to service providers free of charge; thus further enhancing trust and credibility in the EU digital services market on the one hand, and ensuring that preventive blocking mechanisms function in a similar manner across the EU internal market on the other.

As a result, the Centre's concept seems to address two of the main limitations of the system introduced in the DSA for combating the distribution of illegal content – namely too much discretion and a lack of transparency in the decisions taken by service providers. At the same time however, as the draft regulation is only at an early legislative stage, it is impossible to judge at this stage if and when the legal solutions proposed therein will come into force. Moreover, the CSAM Regulation only concerns the narrow area of identification and blocking of paedophile content, so it does not affect the application of identical mechanisms concerning other types of illegal material – blocked either on the basis of the DSA or other specific provisions (e.g. the Terrorist Content Regulation).

An alternative to the model proposed in the draft CSAM Regulation is the mechanism introduced in CDSM Directive.[28] This act essentially modifies and extends the liability framework established in the e-Commerce Directive (and currently also in the DSA) by requiring service providers to "make, in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works and other subject matter."[29] In effect, the CDSM Directive establishes more far-reaching content filtering obligations than those traditionally associated with the *notice and takedown* model (Romero-Moreno, 2020). The discussion concerning the proportionality of the solution adopted in the CDSM Directive –

[27] Recital (23) of the Regulation.

[28] See n 12 above.

[29] Art. 17(4)(b) CDSM Directive.

and its compatibility with EU law – lasted for years[30] and eventually led to a complaint to the CJEU, in which Poland sought the annulment of the preventive content filtering mechanisms established in the act.

In its 2022 judgment, the Court – accepting the position of the Advocate General[31] – held that an obligation to remove specific content does not lead to the establishment of a general obligation to monitor, and therefore does not infringe the e-Commerce Directive (see earlier comments on Article 15(1) of the e-Commerce Directive and the current Article 8 of the DSA).[32] In principle, it reiterated and extended its position previously expressed in the Glawischnig-Piesczek case by pointing out that the identification of infringing works does not require a service provider to "independently assess" the content of those works, and that it is sufficient to rely on the identifying information provided by copyright owners.[33] Leaving aside disputes over the definition of the term "independent assessment" – especially in context of how the self-learning algorithms commonly used in the mass data market operate – the mechanism introduced for identifying illegal content is also controversial. The notice and takedown model was developed to protect the rights of users who were harmed by a particular publication that targeted their legally-protected interest. This model shifted the burden of proof to the party that reported the infringement. Although this principle has been maintained in the CDSM Directive, given the massive nature of copyright infringement and the negative financial consequences of failing to comply with the requirements of the Directive, there is a risk that service providers will act passively; i.e. not carry out sufficient verification, but rather block all material designated by rights management organisations (Spindler, 2019, 355). This may lead to a further reduction in the transparency of the whole process, with one private actor (the service provider) following instructions from a number of other private actors (copyright owners or – more often – collecting societies), without sufficient oversight by public authorities.[34]

It is clear that the content filtering mechanisms laid down in the draft CSAM Regulation and the CDSM Directive serve different purposes and correspond to infringements of a different nature and also of different gravity. According to the CJEU's interpretation of the principle of proportionality,[35] this may justify the adoption of more far-reaching measures in the case of the dissemination of paedophilic content than those applied to copyright protection. However, it appears that the measures introduced by the CSAM Regulation – such as the establishment of a reference model for content filtering or the creation of a common EU register of blocked content – are mechanisms that could be equally effective in combating other instances of the distribution of illegal content online. Moreover, their implementation would help to remove a number of doubts regarding the compatibility of proactive content filtering measures with EU law, including those relating to the adequacy of the oversight mechanisms in place, and would thus help to establish not only a common legal framework, but also the practical application thereof throughout the European Union.

---

30 See e.g. Ferri (2021); Geiger & Jütte (2021); Romero-Moreno (2019).

31 See in particular: Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Case C-401/19, EU:C:2021:613, para. 115.

32 Judgment of 26 April 2022, *Republic of Poland v European Parliament and Council of the European Union,* C-401/19, EU:C:2022:297.

33 Ibid., para. 90.

34 For a broader discussion on the Court of Justice's ruling on the CDSM Directive, see (Jütte, 2022).

35 For a broader discussion on the Court of Justice's ruling on the CDSM Directive, see (Jütte, 2022).

5 Summary

The dynamic changes in the global digital market have highlighted not only the opportunities but also the challenges faced by service providers, users and legislators alike. Owing to the mass nature of many services – in particular electronic media and social networks – nowadays, for a large proportion of users, digital services are their main source of information, replacing the need for traditional media (Westerman et al., 2014). However, given the technical possibilities available, this increasing use of digital services has also revealed a number of threats to the protection of individual rights, ultimately affecting the opportunity of building a modern information society. Such threats particularly relate to the ease with which untrue or defamatory content can be distributed, and how easily it can gain popularity and reach huge audiences, even in cases where its illegality seems obvious.

The changes introduced by the DSA are the EU legislature's response to these threats, which proposes extending the liability of service providers for illegal speech made by users. The EU legislature thus has taken the view that given the obvious economic asymmetry between the service provider and the user, specific obligations should be imposed on the service provider not only to eliminate illegal content, but to prevent its further dissemination. This leads to the re-shaping of the role of the service provider as no longer being a mere passive intermediary in the transmission of content, but an actor actively influencing the shape of the digital environment in which the user operates.

In this respect, the DSA undoubtedly fits into the regulatory model already introduced in specific acts regulating the functioning of the digital market – such as the CDSM Directive or the Terrorist Content Regulation. While in principle maintaining the prohibition – fundamental to the EU legal model – against imposing a legal obligation on service providers to actively search for illegal content, the DSA nevertheless has introduced a number of incentives for such activities.

Therefore, the DSA moves away from the previously dominant view of the service provider as a 'passive' intermediary in the transmission of information, and towards defining it as an entity obliged to respond to emerging infringements – not only when they are reported but also on its own initiative. As a result, there is a need for much wider use of algorithmic measures, including those involving proactive content monitoring. While the purpose of implementing this type of regulation seems obvious, one should not lose sight of the fact that compliance with the standards of necessity and proportionality requires a precise definition of the type of content that justifies the use of such automated measures. The broad category of "illegal content" seems to be too vague in this regard, and as a result creates the risk of arbitrary decisions. This will force providers of transnational digital services to develop their own content filtering rules in an attempt to reflect the requirements contained in the often divergent regulations in force in individual Member States.

In the course of the work on the DSA, a number of concerns were raised about the appropriateness of positioning service providers as "arbiters" – who in turn themselves create, implement and ultimately evaluate the application of content filtering standards (Jørgensen & Pedersen, 2017). This problem is particularly relevant in the case of upload filters and contributes to the recurrent allegations that the EU is introducing a preventive censorship model. It appears that the EU legislature – while recognising the problem – has not taken sufficient steps to adequately mitigate this risk through the provisions of the DSA.

Only once the DSA is being applied will it be possible to assess the extent to which allegations formulated today have turned out to be accurate. However, it is already clear that there is a need for further discussion on the future regulatory strategy for the application of preventive measures – a discussion that focuses in particular on the development of mechanisms that would limit or eliminate possible abuses arising from excessive content blocking.

References

Bloch-Wehba, H. (2020). Automation in Moderation. Cornell International Law Journal, 53, 42–96. Online: https://ssrn.com/abstract=3521619

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? Ethnicities, 18(3), 297–326. https://doi.org/10.1177/1468796817709846

Cavaliere, P. (2019). Glawischnig-Piesczek v Facebook on the Expanding Scope of Internet Service Providers' Monitoring Obligations (C-18/18 Glawischnig-Piesczek v Facebook Ireland). European Data Protection Law Review, 5(4), 573–578. https://doi.org/10.21552/edpl/2019/4/19

Cox, N. (2014). Delfi AS v Estonia: The Liability of Secondary Internet Publishers for Violation of Reputational Rights under the European Convention on Human Rights. Modern Law Review, 77(4), 619–629. https://doi.org/10.1111/1468-2230.12081

Echikson, W., & Knodt, O. (2018). Germany's NetzDG: A key test for combatting online hate (2018/09). Centre for European Policy Studies. Online: https://cli.re/Bvv1Zx

Ferri, F. (2021). The dark side(s) of the EU Directive on copyright and related rights in the Digital Single Market. China-EU Law Journal, 7, 21–38. https://doi.org/10.1007/s12689-020-00089-5

Fino, A. (2020). Defining Hate Speech. Journal of International Criminal Justice, 18(1), 31–57. https://doi.org/10.1093/jicj/mqaa023

Fuster, G. G., & Jasmontaite, L. (2020). Cybersecurity Regulation in the European Union: The Digital, the Critical and Fundamental Rights. In M. Christen, B. Gordijn, & M. Loi (Eds.), The Ethics of Cybersecurity. The International Library of Ethics, Law and Technology, vol 21 (pp. 97–115). Springer International Publishing. https://doi.org/10.1007/978-3-030-29053-5_5

Geiger, C., & Jütte, B. J. (2021). Platform Liability Under Art. 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match. GRUR International, 70(6), 517–543. https://doi.org/10.1093/grurint/ikab037

Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2), 205395172094323. https://doi.org/10.1177/2053951720943234

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), 205395171989794. https://doi.org/10.1177/2053951719897945

Griffin, R. (2022). New school speech regulation as a regulatory strategy against hate speech on social media: The case of Germany's NetzDG. Telecommunications Policy, 46(9), 102411. https://doi.org/10.1016/j.telpol.2022.102411

Hochmann, T. (2022). Hate speech online: The government as regulator and as speaker. Journal of Media Law, 14(1), 139–158. https://doi.org/10.1080/17577632.2022.2085014

Isaac, A., Kumar, R., & Bhat, A. (2022). Hate Speech Detection Using Machine Learning Techniques. In S. Aurelia, S. S. Hiremath, K. Subramanian, & S. Kr. Biswas (Eds.), Sustainable

Advanced Computing, vol 840 (pp. 125–135). Springer. https://doi.org/10.1007/978-981-16-9012-9_11

Jørgensen, R. F., & Pedersen, A. M. (2017). Online Service Providers as Human Rights Arbiters. In M. Taddeo, & L. Floridi (Eds.), The Responsibilities of Online Service Providers, vol 31 (pp. 179–199). Springer International Publishing. https://doi.org/10.1007/978-3-319-47852-4_10

Julià-Barceló, R., & Koelman, K. J. (2000). Intermediaty Liability. Computer Law & Security Review, 16(4), 231–239. https://doi.org/10.1016/S0267-3649(00)89129-3

Jütte, B. J. (2022). Poland's challenge to Article 17 CDSM Directive fails before the CJEU, but Member States must implement fundamental rights safeguards. Journal of Intellectual Property Law & Practice, 17(9), 693–695. https://doi.org/10.1093/jiplp/jpac076

Keller, D. (2020). Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling. GRUR International, 69(6), 616–623. https://doi.org/10.1093/grurint/ikaa047

Kikarea, E., & Menashe, M. (2019). The global governance of cyberspace: Reimagining private actors' accountability: Introduction. Cambridge International Law Journal, 8(2), 153–170. https://doi.org/10.4337/cilj.2019.02.00

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. Harvard Law Review, 131(6), 1598–1670. Online: https://tinyurl.com/3y5hvrez

Koebler, J., & Cox, J. (2018, August 23). The impossible job: inside Facebook's struggle to moderate two billion people. Vice. Online: https://cli.re/mrDwrA

Kuczerawy, A. (2020). From 'Notice and Takedown' to 'Notice and Stay Down': Risks and Safeguards for Freedom of Expression. In G. Frosio (Ed.), Oxford Handbook of Online Intermediary Liability (pp. 523–543). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198837138.013.27

Lee, H.-E., Ermakova, T., Ververis, V., & Fabian, B. (2020). Detecting child sexual abuse material: A comprehensive survey. Forensic Science International: Digital Investigation, 34, 301022. https://doi.org/10.1016/j.fsidi.2020.301022

Leerssen, P. (2023). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. Computer Law & Security Review, 48, 105790. https://doi.org/10.1016/j.clsr.2023.105790

Lemmens, K. (2022). Freedom of Expression on the Internet after Sanchez v France: How the European Court of Human Rights Accepts Third-Party 'Censorship'. European Convention on Human Rights Law Review, 3(4), 525–550. https://doi.org/10.1163/26663236-bja10046

Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. Big Data & Society, 7(1), 205395172092068. https://doi.org/10.1177/2053951720920686

Mitrakas, A. (2018). The emerging EU framework on cybersecurity certification. Datenschutz Und Datensicherheit – DuD, 42(7), 411–414. https://doi.org/10.1007/s11623-018-0969-2

Molter, S. (2022). Combating hate crime against LGBTIQ* persons. Institute for Social Work and Social Education. Online: https://cli.re/83zrbb

Parsons, C. (2019). The (In)effectiveness of Voluntarily Produced Transparency Reports. Business & Society, 58(1), 103–131. https://doi.org/10.1177/0007650317717957

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate Speech: A Systematized Review. SAGE Open, 10(4), 215824402097302. https://doi.org/10.1177/2158244020973022

Peršak, N. (2022). Criminalising Hate Crime and Hate Speech at EU Level: Extending the List of Eurocrimes Under Article 83(1) TFEU. Criminal Law Forum, 33(2), 85–119. https://doi.org/10.1007/s10609-022-09440-w

Rauchegger, C., & Kuczerawy, A. (2020). Injunctions to Remove Illegal Online Content under

the Ecommerce Directive: Glawischnig-Piesczek. Common Market Law Review, 57(5), 1495–1526. https://doi.org/10.54648/cola2020745

Rojszczak, M. (2023). Gone in 60 Minutes: Distribution of Terrorist Content and Free Speech in the European Union. Democracy and Security, 1–31. https://doi.org/10.1080/17419166.2023.2250731

Romero-Moreno, F. (2020). 'Upload filters' and human rights: Implementing Article 17 of the Directive on Copyright in the Digital Single Market. International Review of Law, Computers & Technology, 34(2), 153–182. https://doi.org/10.1080/13600869.2020.1733760

Romero-Moreno, F. (2019). 'Notice and staydown' and social media: Amending Article 13 of the Proposed Directive on Copyright. International Review of Law, Computers & Technology, 33(2), 187–210. https://doi.org/10.1080/13600869.2018.1475906

Siegel, A. A. (2020). Online Hate Speech. In N. Persily, & J. A. Tucker (Eds.), Social Media and Democracy: The State of the Field, Prospects for Reform (1st ed.) (pp. 56–88). Cambridge University Press. https://doi.org/10.1017/9781108890960

Spano, R. (2017). Intermediary Liability for Online User Comments under the European Convention on Human Rights. Human Rights Law Review, 17(4), 665–679. https://doi.org/10.1093/hrlr/ngx001

Spindler, G. (2017). Responsibility and Liability of Internet Intermediaries: Status Quo in the EU and Potential Reforms. In T.-E. Synodinou, P. Jougleux, C. Markou, & T. Prastitou (Eds.), EU Internet Law (pp. 289–314). Springer International Publishing. https://doi.org/10.1007/978-3-319-64955-9_12

Spindler, G. (2019). The Liability system of Art. 17 DSMD and national implementation. Journal of Intellectual Property, Information Technology and E-Commerce Law, 10(3), 344–374. Online: https://www.jipitec.eu/issues/jipitec-10-3-2019/5041

Teršek, A. (2020). Common and Comprehensive European Definition of Hate-Speech Alternative Proposal. Open Political Science, 3(1), 213–219. https://doi.org/10.1515/openps-2020-0019

Wang, J. (2018). Notice-and-Takedown Procedures in the US, the EU and China. In J. Wang, Regulating Hosting ISPs' Responsibilities for Copyright Infringement (pp. 141–178). Springer. https://doi.org/10.1007/978-981-10-8351-8_5

Westerman, D., Spence, P. R., & Van Der Heide, B. (2014). Social Media as Information Source: Recency of Updates and Credibility of Information. Journal of Computer-Mediated Communication, 19(2), 171–183. https://doi.org/10.1111/jcc4.12041

Wilman, F. G. (2022). Two emerging principles of EU internet law: A comparative analysis of the prohibitions of general data retention and general monitoring obligations. Computer Law & Security Review, 46, 105728. https://doi.org/10.1016/j.clsr.2022.105728

Wu, F. T. (2013). Collateral Censorship and the Limits of Intermediary Immunity. Notre Dame Law Review, 87(1), 293–349. Online: https://scholarship.law.nd.edu/ndlr/vol87/iss1/6/