Available online at www.sciencedirect.com

ScienceDirect

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

Computer Law & Security Review

Online content filtering in EU law – A coherent framework or jigsaw puzzle?

Marcin Rojszczak[1]

Faculty of Administration and Social Sciences, Warsaw University of Technology, Warsaw, Poland

ARTICLE INFO

Keywords:
Content filtering
Online censorship
Prior restraint
Notice and takedown

ABSTRACT

With the spread of global digital services, the need to establish effective barriers to the dissemination of illegal content has also grown. Online services, instead of supporting the building of social relationships and allowing the free exchange of ideas, are increasingly becoming platforms for spreading hate speech or promoting extremist behaviour. The commitment to respect fundamental rights on which the Union has been built also requires those rights to be protected in cyberspace. Increasingly, one measure implemented to achieve this goal is filtering or blocking illegal content.

In recent years, as a result of both the jurisprudence of European courts and the activity of the EU legislature, content filtering measures are increasingly used in an automatic and often also preventive manner. The freedom to use them on the part of digital service providers raises obvious concerns about compliance with human rights standards, often leading to allegations of implementing a new form of "digital censorship".

Assuming that content filtering will be a measure that will be increasingly used in an automatic manner, it is particularly important to establish adequate standards of legal safeguards to protect against the risk of their abuse.

The purpose of this article is to explain why the regulations currently being introduced do not create a coherent regulatory model and *de facto* hamper the effective protection of end-user rights and public security objectives, by introducing a series of often overlapping legal requirements. In this respect, the mosaic of various regulations does not facilitate the definition of a coherent standard of legal safeguards, which in turn delimits the boundaries of the application of automatic content filtering measures.

1. Introduction

Cyberspace is commonly equated with freedom of communication. It is presented as a borderless space where anyone can express their views and opinions, possibly to millions of other users. As a result, attempts to discuss the limits on freedom of speech in cyberspace are often met with criticism, and in extreme cases even lead to social unrest.

However, the mass use of online services has also made cyberspace a place for disseminating and promoting violent, defamatory and exclusionary content, as well as inciting and promoting crime, including terrorism.[1] There is no doubt that the fight against terrorism is an important and ongoing task

E-mail address: marcin.rojszczak@pw.edu.pl
[1] ORCID ID: 0000-0003-2037-4301.

[1] ECtHR, Appl. 64569/09, Delfi AS v. Estonia (2015), para. 110.

https://doi.org/10.1016/j.clsr.2022.105739

for public authorities, and it would be difficult to justify passivity on the part of governments in preventing the dissemination of radical content based on the idea that it is protected by freedom of speech in cyberspace. The fight against terrorism is only one case where a state response seems necessary in order to protect general security. The spread of child sexual abuse content is an equally serious problem.[2] While terrorist threats are defined and perceived differently in different parts of the world based on differences in worldviews or cultures,[3] there seems to be a global, universal recognition of the need to firmly tackle child sexual abuse.

Certainly, the modern Internet is a place where its apparent anonymity creates a huge scope for the promotion of hate and xenophobic speech. The growing wave of aggressive content has become a major reason leading social media providers to implement their own mechanisms for controlling and moderating content shared by users.[4] Although the need to take such steps seems obvious, it also leads to questions about the scope of permissible interference – in particular the legality of using certain preventive control measures in a democratic state.[5] These are also concerns about the adequacy of measures taken, their conformity with the legal model in force, as well as whether they ensure respect for the principles of proportionality and strict necessity. Voluntary controls applied by digital service providers and based on *soft law* mechanisms have raised doubts about their respect for freedom of speech, intensifying discussion on the need for the EU legislature to take more comprehensive action.

For several years now, the issue of developing consistent rules for the preventive filtering of online content has been the subject of legislative activity, as well as being an area of interpretation by the European courts.[6] Although the actions undertaken can hardly be regarded as comprehensive, the very fact that they have been carried out has understandably aroused the interest of the public.

The purpose of this article is to discuss the existing EU legislation that *per se* introduces preventive content filtering measures or creates a framework for establishing such mechanisms for digital services. The research scope of this paper in particular covers provisions that place obligations on service providers to provide automatic content filtering, often conducted in a preventive manner.

The analysis will focus on three main areas: (i) protection against defamation and hate speech; (ii) protection of intellectual property rights; and (iii) measures related to the protection of public security. Discussion of the EU legal model will be preceded by terminological considerations in order to introduce some systematisation and explain what content filtering is and how it is implemented in practice. Such an introduction will facilitate the subsequent analysis of the various mechanisms present in EU law and assessment of the extent to which their introduction interferes with fundamental rights.

The article argues that the evolution of the content filtering law does not lead to the establishment of a coherent regulatory framework. In turn, the introduction of a clear regulatory model is necessary both in order to protect the rights of individuals as well as the need for clarity of the rules applicable to digital service providers. The further introduction of fragmented and partially inconsistent content filtering rules will eventually create a barrier to the development of the digital services market in the EU.

2. Different dimensions of content control

Any discussion about online content filtering measures should be preceded by an explanation of the relationship between content control and censorship. Such an introduction is particularly important given the frequent (and repeated) complaints about successive content filtering mechanisms introduced in the EU.[7] Thus the meaning of the prohibition of censorship, introduced both in the constitutional provisions of Member States as well as in the jurisprudence of the European courts, requires clarification.

In general, any control of publications – including online content – can be divided into two categories: preventive and reactive (*ex post*). Reactive control is usually carried out by a court (or other competent authority) and is related to the adjudication of a specific case. As a rule, it can be carried out on the initiative of both an individual (as a legal remedy) or a public entity (acting in the public interest). In either case, reactive control concerns, in principle, content that has already been published. Therefore, its application is seen as a more proportionate interference with the rights to free speech and information than preventive control.

Preventive control, on the other hand, aims to prevent possible infringements resulting from a publication. In this regard, the term *censorship* is also often used – most frequently as a synonym for *institutional censorship*, i.e. making a publication contingent on the prior consent of a designated public authority. This measure usually takes the form of *general censorship* –

[2] Maxwell Taylor and Ethel Quayle, *Child Pornography: An Internet Crime* (Brunner-Routledge 2003); Henry Hillman, Christopher Hooper and Kim-Kwang Raymond Choo, 'Online Child Exploitation: Challenges and Future Research Directions' (2014) 30 Computer Law & Security Review 687.

[3] *See e.g.* Boaz Ganor, 'Defining Terrorism: Is One Man's Terrorist Another Man's Freedom Fighter?' (2002) 3 Police Practice and Research 287; JM Sorel, 'Some Questions About the Definition of Terrorism and the Fight Against Its Financing' (2003) 14 European Journal of International Law 365.

[4] Jennifer Cobbe, 'Algorithmic Censorship by Social Platforms: Power and Resistance' [2020] Philosophy & Technology <http://link.springer.com/10.1007/s13347-020-00429-0> accessed 14 September 2021.

[5] For more on the different approaches to moderating social media content, see: Botambu Collins and others, 'Trends in Combating Fake News on Social Media – a Survey' (2021) 5 Journal of Information and Telecommunication 247.

[6] See Comparative study on blocking, filtering and take-down of illegal internet content, Council of Europe (2017), <https://cli.re/9DBDZ2> accessed 24 September 2021.

[7] See e.g. Kalev Leetaru, 'The EU Will Be The End Of Free Speech Online', Forbes (6 July 2019), <https://cli.re/EoZMyy> accessed 24 September 2021; Eva Simon, 'Say No To Online Censorship in Europe!', Liberties (28 June 2018), <https://cli.re/MD4ejB> accessed 24 September 2021; Cory Doctorow, 'EU Internet Censorship Will Censor the Whole World's Internet', Electronic Frontier Foundation (9 October 2018), <https://cli.re/ZWKxXJ> accessed 24 September 2021.

that is, covering all publications made available in all media. In fact, for the classification of censorship as institutional it is irrelevant whether the actual control is carried out by a public entity: control carried out by a private entity – but according to the instructions and under the supervision of a public authority – can have the same effect as institutional control. In either case, the distinguishing feature that makes it possible to classify a given measure as a mechanism of preventive control is the purpose of its application – which, insofar as concerns general censorship, often encompasses areas not considered as part of a democratic state (e.g. political or moral control[8]).

Preventive control, however, should not be equated exclusively with institutional censorship. In democratic states, preventive control is also exercised by courts within the limits set by law. It can therefore be incidental censorship (which is the opposite of general censorship). Even in the United States, where the First Amendment to the Constitution specifically protects freedom of speech, the US Supreme Court has created a precedent by allowing preventive control (known as *prior restraint*[9]) to be applied in exceptional cases.[10] These include the protection of national security[11] and cases involving the publication of state secrets.[12] In its rulings, the Court has also stressed that not every publication should be covered by the First Amendment, as the Amendment only covers content having a "literary, artistic, political or scientific value".[13] As a result, publications that focus on promoting violence[14] or obscene content[15] are not protected by the US Constitution.

It should be noted that institutional censorship need not always be of a general nature. An example is measures applied to specific categories of publications, such as foreign or foreign-language publications, whose dissemination may be subject to prior administrative control.[16]

The prohibition of censorship has also been *explicitly* introduced into the basic law many European countries – as a norm reinforcing the importance of freedom of expression.[17] Nevertheless, this prohibition applies only to institutional censorship. Such a conclusion can be drawn from, *inter alia*, a judgment of the German Constitutional Court in which it stated that the prohibition of censorship under Article 5(1) of the Constitution should be understood as blocking the possibility of making the publication or distribution of a work conditional on the prior approval of its content by a public authority. In this respect, the Court stressed that "the very existence of such a control and approval procedure paralyses intellectual life".[18] Hence, the prohibition of institutional censorship as defined in Germany's Constitution cannot be affected even by the application of the "general laws" exception.[19] A similar position was taken by the Polish Constitutional Tribunal, which explained that the prohibition of preventive censorship arising from Article 64(1) of the Constitution should be understood as obstacle to "making the publication or broadcast of a particular message dependent on the prior consent of a public authority".[20] The prohibition of censorship is also present in the constitutions of Spain, Portugal and Italy.[21] In many cases though, the law allows the use of preventive censorship, for example in the case of protecting "the right to honour, to privacy, to personal reputation and to the protection of youth and childhood" (Spain),[22] or any "offence against the Christian or any other known religion" (Greece).[23] However, in these cases as well preventive control must be implemented on the basis and within the limits of a court order.

A prohibition of censorship was not explicitly introduced into the European Convention on Human Rights. Nevertheless, when examining complaints about the violation of freedom of expression, the ECtHR has repeatedly commented on the role and importance of freedom of expression for the development of democratic society,[24] as well as the status of the media as a "public watchdog" monitoring the activities of those in power.[25] At the same time, the Court noted that "prior restraints are not necessarily incompatible with the Convention as a matter of principle".[26] This pertains to the application of institutional control measures, but not those of a gen-

[8] *See* the opinion of AG Henrik Saugmandsgaard Øe in case C-401/19, Poland v Parliament and Council, EU:C:2021:613 para. 77.

[9] It should be recalled that censorship and prior restraint are not the same thing. As Edwin S. Newman points out, "the difference between censorship and prior restraint is not easy to define (...) Analytically, however, censorship is aimed against the transmission of an idea, thought or expression, while prior restraint is aimed against the particular method of transmission – rather than specifically against the idea." Newman, Edwin S., Editor. Freedom Reader: A Collection of Materials on Civil Rights and Civil Liberties in America. New York, Oceana Publications.

[10] Thomas I Emerson, 'The Doctrine of Prior Restraint' (1955) 20 Law and Contemporary Problems 648; John Calvin Jeffries, 'Rethinking Prior Restraint' (1983) 92 Yale Law Journal 409. *See generally* Emerson; Jeffries.

[11] Defined in a very narrow sense. *See*, e.g., US Supreme Court, Near v Minnesota, 283 US 716 (1931). See also Geoffrey R. Stone, Free Speech and National Security

[12] US District Court for the Western District of Wisconsin, United States v Progressive Incorporated and ors, 486 F Supp 5 (WD Wis 1979); see also the reasoning presented by the Court in the Pentagon Papers case – US Supreme Court, New York Times Company v United States, 403 US 713 (1971).

[13] US Supreme Court, Miller v California, 413 US 15 (1973).

[14] US Supreme Court, Brandenburg v Ohio, 395 US 444 (1969).

[15] US Supreme Court, Roth v United States, 354 US 476 (1957).

[16] ECtHR, Appl. 39288/98, Association Ekin v. France (2001), para. 58.

[17] Eleni Polymenopoulou, 'Censorship' in Rainer Grote, Frauke Lachenmann and Rudiger Wolfrum (eds), *Max Planck Encyclopedia of Comparative Constitutional Law* (Oxford University Press).

[18] Bundesverfassungsgericht, 25 April 25 1972, 1 BvL 13/67 (BVerfGE 33, 52), para. 76.

[19] *Ibid*. According to Art. 5(2) of the Basic Law, freedom of speech "shall find [its] limit in the provisions of general laws". When referring to this limitation, the German Constitutional Court stated, that "the ban on censorship is intended to banish the typical dangers of such a preventive control. That is why there must be no exception to the ban on censorship, not even through "general laws" according to Article 5, Paragraph 2 of the Basic Law."

[20] Constitutional Court, 9 November 2010, K 13/07, para. III.2.

[21] *See* Art. 20(2) of the Spanish Constitution; Art. 37(2) of the Portuguese Constitution; Art. 21 of the Italian Constitution

[22] Article 20(4) of the Spanish Constitution.

[23] Article 14(2)(a) of the Greek Constitution.

[24] ECtHR, Appl. 3002/03 and 23676/03, Times Newspapers Ltd v United Kingdom (2009), para. 27.

[25] ECtHR, Appl. 17488/90, Goodwin v United Kingdom (1996), para. 39.

[26] ECtHR, Appl. 3111/10, Ahmet Yıldırım v. Turkey (2012), para. 64.

eral nature.[27] The Court emphasised that the application of preventive control requires the establishment of appropriate legal safeguards "ensuring both tight control over the scope of bans and effective judicial review to prevent any abuse of power".[28] At the same time, the domestic court's assessment aimed at balancing the competing interests cannot be based on too much discretion and must be grounded on precise legal rules. Applying these criteria, the ECtHR concluded that the Turkish court's approval of the blocking of access to Google's services on the grounds of an alleged threat to national security failed to meet these criteria, as the national court did not examine whether this was the least intrusive measure that could be taken in the given situation.[29]

Therefore, while ECtHR case law does not directly prohibit preventive control, the model for carrying out such control must be based on strict judicial supervision and treated as an exception to the general prohibition of unlawful interference into freedom of expression. This conclusion was also confirmed by the position of the Court in *Cumpănă and Mazăre v. Romania*, according to which "prior restraints on the activities of journalists call for the most careful scrutiny on its part and are justified only in exceptional circumstances".[30]

The Council of Europe acquis also addresses the unacceptability of institutional censorship of online content. According to the *Declaration on freedom of communication on the Internet*, adopted by the Committee of Ministers of the Council of Europe, public authorities should refrain from applying general measures of blocking and filtering content that prevent public access to it.[31] Importantly, exceptions are allowed for the protection of minorities and the blocking or removal of specific content, the unlawfulness of which has been established by a final decision of the competent national authority.

A separate issue is determining whether the content of constitutional norms and international obligations introducing the prohibition of unlawful interference in freedom of speech also implies positive obligations on the part of states in shaping their national law to prevent the emergence of mechanisms of preventive censorship used by private entities. In recent years, a lot of attention has been paid to actions taken by major social networking platforms in the fight against *fake news* and hate speech.[32] Service providers, based on their terms of service, have introduced their own content filtering measures – often not very transparent and not subject to external oversight.[33] If the ban on censorship were also considered to entail positive obligations, it would be reasonable to propose the introduction of legislation subjecting the content control mechanisms voluntarily implemented by service providers to detailed regulation.

3. Technical aspects of automatic content filtering

In any discussion on the filtering of online content, a great deal of public attention is focused on the issue of the mechanisms used by service providers. However, it should be borne in mind that information transmitted over the Internet may be altered at any stage of its transfer. Access to it may therefore be filtered either at source (i.e. by the content provider); by the Internet service providers (ISPs); by any of the entities it uses (infrastructure providers); as well as directly on the user's end device.

In practice, the most effective way to filter online content is to block information at source. At the same time, such a measure obviously requires the cooperation of the content provider, which is not always possible, especially considering the cross-border nature of the services provided. Data filtering by the service provider potentially restricts access to content for all users, which is also not always a desirable outcome. An example is the exercise of the right to be forgotten under Regulation 2016/679, which according to the CJEU case law should lead to content blocking, but only for EU users.[34] As a result, the content provider may be subject to different, potentially conflicting, court orders aimed at preventing access to the same content by different audiences (e.g. users from different countries). Examples of content filtering at source are the mechanisms used by file sharing and streaming providers (e.g. YouTube).

Blocking access to information by infrastructure operators generally consists of preventing the use of certain Internet resources. Therefore, it is not *strictly* content filtering, but disallowing access to certain resources. In this way telecom operators can also protect their users against certain types of malware by detecting or preventing communications from infected computers from being transmitted to hackers' servers (the so-called C&C servers).[35] Thus this technique may also be used to achieve cybersecurity goals in a way that is not related to content control and has no impact on freedom of speech.

Nevertheless, it should be noted that leaving the provision of content control mechanisms solely to infrastructure providers is both impractical and ineffective. Firstly, not every data stream can be analysed in real time during transmission (e.g. due to the encryption used). Secondly, as a result of the very nature of the Internet structure, information may be

[27] ECtHR, Appl. 39288/98, Association Ekin v. France (2001), para. 58.

[28] ECtHR, Appl. 33014/05, Editorial Board of Pravoye Delo and Shtekel v. Ukraine (2011), para. 55.

[29] ECtHR, Appl. 3111/10, Ahmet Yıldırım v. Turkey (2012), para. 68.

[30] ECtHR, Appl. 33348/96, Cumpănă and Mazăre v. Romania (2004), para. 118.

[31] *See* Principle No. 3 in the Declaration on freedom of communication on the Internet adopted by the Committee of Ministers on 28 May 2003, <https://cli.re/mrE5Vb> accessed 24 September 2021.

[32] Collins and others (n 4); Cobbe (n 3); Catherine O'Regan, 'Hate Speech Online: An (Intractable) Contemporary Challenge?' (2018) 71 Current Legal Problems 403.

[33] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598. The practice is therefore just as often subject to criticism, see e.g. ibid.

[34] *See also* Jure Globocnik, 'The Right to Be Forgotten Is Taking Shape: CJEU Judgments in GC and Others (C-136/17) and Google v CNIL (C-507/17)' (2020) 69 GRUR International 380.CJEU, Case C-507/17, Google LLC v CNIL (2019) EU:C:2019:772, para 73.

[35] Futai Zou and others, 'Detecting Malware Based on DNS Graph Mining' (2015) 2015 International Journal of Distributed Sensor Networks 1.

transmitted via alternative routes, bypassing providers who use data filtering mechanisms. Thirdly, while an infrastructure provider may simply block access to a specific server or prevent a user from downloading a specific file, analysing the file's content is a much more intensive task in computational terms. Given the volume of information exchanged on the Internet, it is in practice impossible.

Having said that, this method is relatively effective for blocking access to an entire service or all data originating from a specific source (i.e. a specific online service). It is therefore an adequate solution when the service provider does not wish to cooperate with the authorities (e.g. is located in a foreign jurisdiction) or when the entire service offered by it violates national law. Therefore, this mechanism has been applied in some EU countries as a tool to prevent the use of online gambling services provided without the licence required under national law. Such a system has already been implemented in, e.g. Poland, where the Ministry of Finance has been authorised to maintain a special register of websites to which access must be blocked by telecommunication companies.[36] As of September 2021, more than 15,000 Internet domains have been listed in the register. Nonetheless, the evidence from Poland also illustrates the weaknesses of this approach. Although the register has been functioning for five years, one can still easily use the services noted down on the register via a significant number of Internet providers. The reason for this is the way the blockade is implemented, as it is based on introducing changes to the DNS[37] registries kept by telecommunication operators. A user does not have to use the DNS service of their operator and is free in this respect. They can therefore, for example, use a service directly offered by other entities, including foreign ones (e.g. DNS servers provided by Google). In such a case, any restrictions imposed at the national level will not apply. This situation leads to the inevitable conclusion that basing the rules for filtering data and services on infrastructure providers requires the introduction of a transnational solution whereby entities offering their services from different countries would be subject to the same requirements and restrictions.

A third place to impose restrictions on access to information is on the user device itself (such as a laptop or mobile device). Such measures are often implemented by businesses to restrict employees' access to non-work related content using a corporate network. The market also offers a number of tools for the so-called 'parental control', which operates in a similar way. Nevertheless, when using this type of technology there is no direct risk of introducing preventive censorship, because *a priori* the measures used are not of a general nature and are not implemented or supervised by public authorities. This assessment however does not change the fact that this technology can potentially be a very effective surveillance tool. This is evidenced by a recent statement from Apple, which announced its intention to add functionality to its mobile devices capable of identifying and reporting child pornography.[38] Such control would be carried out automatically, without the user's interference, on the basis of the terms of use accepted by the user during the device start-up (i.e. the installation of system software).[39] Leaving aside the legitimate goal behind the introduction of this measure, one cannot fail to notice the dangers resulting from the spread of such a practice – which can be used just as effectively to filter any kind of information. For example, Xiaomi devices have built-in mechanisms for filtering certain content designated by the Chinese government.[40] Devices offered on the EU market have this function blocked, but it can be re-activated by the manufacturer at any time. This problem should increase the vigilance of legislators, especially taking into account the fact that a significant number of mobile technologies – hugely popular amongst consumers – were designed and manufactured in non-democratic countries.

In practice, the key issue regarding the effectiveness of automatic content filtering systems is the mechanism used to identify prohibited content. In the case of file inspection, it may be sufficient to use the so-called *hashes* (often called digital fingerprints). However, this solution is ineffective for multimedia files, which can be processed and modified to make identification more difficult. In this case, more sophisticated algorithms, based on machine learning systems, are required. For example, they can ensure that a video sharing service provider blocks access to content with a particularly violent scene and all subsequent files resulting from the processing of the original recording. However, the use of machine learning systems requires adequate capital, competence, and organisational resources. The effectiveness of such systems depends not so much on the initial programming as on the painstaking process of training (teaching) the algorithms to recognise content that violates the rules that have been put in place.[41] Smaller service providers will therefore not be able to ensure proper implementation of this type of technology. Even in the case of large technology companies it is impossible to avoid errors in automated data processing. Taking into account the huge amount of processed information, even a small error – a minuscule percentage – can lead to blocking the publication of thousands of recordings per day. The media often reports

[36] Article 15f(5) of the Gambling Act of 19 November 2009, OJ of 2020 item 2094.

[37] The Domain Name System (DNS) service makes it possible to change a domain name to the IP address of the server on which the requested website is located. The modification of information in the ISP's DNS registry makes it possible to redirect users who wish to access a specific website (e.g. where gambling is offered in violation of the law) to another website designated by the authority (in the case of Poland, the information portal of the Ministry of Finance). In reality, therefore, modifying DNS data is neither filtering information nor strictly blocking it.

[38] Jack Nicas, 'Apple's iPhones Will Include New Tools to Flag Child Sexual Abuse', New York Times (18 August 2021), <https://cli.re/aD51WY> accessed 24 September 2021.

[39] In the same way, Apple has also launched content filtering on its devices used in China. According to a CitizenLab report, Apple products use filtering rules that cover more than 1,100 keywords – most of which relate to politics, human rights, forms of government, and religion. Jeffrey Knockel and Lotus Ruan, 'Engrave Danger: An Analysis of Apple Engraving Censorship across Six Regions', CitizenLab (18 August 2021), <https://cli.re/Ww3m4v> accessed 24 September 2021.

[40] Gareth Corfield, 'Lithuania tells its citizens to throw Xiaomi mobile devices in the bin', The Register (22 September 2021), <https://cli.re/QkoRD4> accessed 24 September 2021.

[41] Cobbe (n 3).

cases of social networking sites blocking content that was deemed offensive or otherwise violated the rules – but which was in fact misinterpreted by the algorithms.[42] Therefore, service providers usually offer the possibility to file a complaint about the blocking of particular content, which initiates a procedure of manual analysis of the case.

Automatic content filtering is usually defined as a preventive measure. In this case, the purpose of its use is to verify unpublished content in order to identify violations of the law (so-called "upload control"). On the other hand, blocking or removing content at the request of interested parties (for example, a notice from copyright owners) usually requires human intervention and is not fully automatic. Despite this, machine learning systems are nonetheless being implemented more and more often. They are used to identify publications that are similar or identical to those previously reported as infringing the law. As a result, the same algorithms that perform upload control can also be used to identify illegal content that has already been published online (so-called "on-going control").[43] Their inherent feature is the incremental effect,[44] as a result of which the same algorithm examining the same input data at different times may come up with a different result due to the knowledge it has acquired in the interim. It should be kept in mind however that the incremental effect not only increases the effectiveness of censorship mechanisms, but at the same time creates a risk of propagating and multiplying an error previously made.

4. EU legal framework of digital services provider's liability

The issue of the liability of service providers (so-called intermediaries' liability) for shared and disseminated online content has been regulated in Directive 2000/31 (the e-Commerce Directive, ECD),[45] which has been in force for over 20 years.

The act contains three fundamental and interrelated provisions – concerning mere conduit (Art. 12); caching (Art. 13); and hosting (Art. 14) – with no obligation to monitor communications (Art. 15). These regulations define the conditions which, when fulfilled, exclude the liability of the provider for the content made available.

In accordance with the mere conduit and caching principles, the service provider is not responsible for the transmitted data unless it is the originator of the transmission or affects the recipient's choice or the content of the transmission (in particular if it modifies the data). Both regulations concern the transmission of information, not its storage, and therefore they are of less importance for the automatic filtering measures analysed in this paper.

The act also defines the conditions under which a service provider may be exempted from liability for content stored at the request of users (Art. 14(1) of the ECD). The first is lack of knowledge about the illegal nature of the content. The second is immediate action (blocking or removal of the information) if reliable information indicating its illegal nature is obtained (the so-called *notice and takedown* model). It is irrelevant how the service provider obtains information about the illegal nature of the content – in particular whether it establishes the information by itself or the information is passed on to it by the victim or a third party. At the same time however, this liability regime does not expressly include an obligation to remove other content that is identical to that challenged or to ensure that similar content is also removed in the future.

As a result, on the one hand adoption of the concept of the *extended* effect of court injunctions – also referred to as the "*notice and staydown*" model – would lead to the establishment of a preventive censorship mechanism (also in this case applied on an individual basis). On the other hand, a finding that a court decision could not affect the ability of other users to publish content containing the same unlawful information would *de facto* deprive the original decision of its effect, and the injured party would be faced with the necessity of bringing subsequent lawsuits in the future to remove information whose unlawfulness had already been established.

Under EU law, the above problem has also been addressed by another standard, defined in Article 15(1) of the e-Commerce Directive, which prohibits Member States from imposing a general obligation to monitor information or to "actively seek facts or circumstances indicating illegal activity".

This prohibition concerns only the imposition of a general oversight obligation on service providers – without affecting the possibility of establishing specific content control requirements. Its establishment does not have the effect of abolishing the liability of suppliers under national law for the consequences of actions or omissions relating to the detection or prevention of unlawful activity.

The e-commerce directive was adopted at a time when the global digital services known today – in particular social networks or media sharing services – did not exist. The development of electronic communications services has not only made it easier for millions of users to communicate freely, but also made it possible for each of them to become an author publishing content that interests them and gaining an audience comparable to traditional media.

These progressive technological changes ever more often raised numerous interpretative doubts, in particular regarding the limits of liability of service providers as well as the possibility (or obligation) of using ever more modern systems for analysing and moderating content published by users. As a result of the jurisprudence of both the Court of Justice and the European Court of Human Rights, some of these ambiguities have been clarified. At the same time however, the EU

[42] Tommi Gröndahl and others, 'All You Need Is "Love": Evading Hate Speech Detection', *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (ACM 2018) <https://dl.acm.org/doi/10.1145/3270101.3270103> accessed 26 September 2021.

[43] More in-depth analysis of different technologies used for content filtering is discussed in: Giovanni Sartor and Andrea Loreggia, 'The impact of algorithms for online content filtering or moderation', European Parliament (2020), <https://cli.re/wrv8dN> accessed on 24 September 2021.

[44] More about incremental learning in: Christophe G Giraud-Carrier, 'A Note on the Utility of Incremental Learning' (2000) 13 AI Commun. 215.

[45] Directive 2000/31 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'), OJ of 2000 L 178/1.

legislator considered it appropriate to adopt a number of new regulations which, often constituting a *lex specialis* for the provisions of the Directive, supplemented the "notice and take-down'' model introduced in it with more proactive content filtering obligations. The following sections discuss the most important areas related to the protection of individual rights or general security objectives in which such measures have been introduced.

This prohibition concerns only the imposition of a general oversight obligation on service providers – without affecting the possibility of establishing content control requirements under specific circumstances. What's more, this provision does not limit the liability of intermediaries under national law for the consequences of actions or omissions relating to the detection or prevention of unlawful activities.

The e-Commerce Directive was adopted at a time when the global digital services known today – in particular social networks or media sharing services - did not exist. The development of electronic communications services has not only made it easier for millions of users to communicate freely but also made it possible for each of them to become an author publishing content of their choice and gaining an audience comparable to traditional media.

These progressive technological changes ever more often led to numerous ambiguities, in particular regarding the limits of liability of service providers as well as the possibility (or obligation) of using increasingly improved systems for analysing and moderating content published by users. As a result of the jurisprudence of both the Court of Justice and the European Court of Human Rights, some of these doubts have been cleared. At the same time, however, the EU legislature considered it appropriate to adopt a number of new regulations, which, often constituting a *lex specialis* for the provisions of the Directive, supplemented the "notice and take-down" model introduced in it with more proactive content filtering obligations. The following sections discuss the most important areas related to the protection of rights of individuals and general security objectives in which such specific measures have been introduced.

5. Recent EU developments regarding automatic content filtering

5.1. *Responding to hate speech and defamation*

The protection of an individual against defamatory information is usually carried out through the courts. Depending on the legal measures adopted, it may be conducted using either criminal or civil procedures, or both. In all cases, however it is the court's ruling that is decisive in confirming that violation has occurred and in obliging certain entities (e.g. digital service providers) to remove the defamatory content.

There is therefore no doubt that such cases should qualify as *ex post* individual censorship. It is not preventive because it concerns content that has already been made available. In view of the ease with which information, including false information, is disseminated in online services, the question has arisen in the case law of national courts as to whether, based on the existing legal framework, it is possible to issue orders for the removal of certain information and all subsequent information having the same content as that considered defamatory. In other words, is it permissible to oblige a service provider to remove or block not only the specific content identified by a claimant (e.g. a file, a video, etc.) but also all repetitions of that information reproduced by other users of the same online service? Finally, may an injunction also be used in relation to future content that, being replications of the same unlawful information, should also be automatically removed (or blocked)?

The above ambiguities resulted in a number of doubts as to the application of the ECD provisions, in particular regarding the obligation for service providers to counteract repeated infringements of the same type. Already in the *L'Oréal SA* case the Court of Justice stated that it was not permissible to impose an obligation on an online service provider to actively monitor all the data transmitted by each user in order to "prevent any future infringement of intellectual property rights via that provider's website".[46]

Moreover, in the *Scarlet Extended SA* case he Court explained that "preventive monitoring of this kind would thus require active observation of all electronic communications (…) and, consequently, would encompass all information to be transmitted and all customers using that network".[47] According to the latter interpretation, prohibiting the verification of information in a general manner, as provided for in the e-Commerce Directive, must be interpreted as referring to the scope of the information processed and not to the volume of information sought. Thus, even an attempt to identify or block a small, clearly defined piece of information – by searching all available data – must be considered to be conducted in a *general* manner and thus in violation of EU law. This interpretation is also supported by the *Mc Fadden v Sony* judgment, in which the Court held that it is impermissible to require a service provider to actively search for and block a single song within the services provided.[48]

It should be noted that the line of jurisprudence presented above has been developed as a result of the Court of Justice's consideration of intellectual property cases. However, in the *Glawischnig-Piesczek v Facebook* case, the Court faced the problem of applying a standard it had developed in disputes concerning the protection of personal rights, including the right to dignity.

The case raised doubts as to whether the domestic court order "may also be extended to statements with identical wording and/or having equivalent content".[49] The CJEU shared the national court's view that, in the case of defamatory or libellous content, the unlawfulness of the content is not necessarily attributable to the use of certain words or phrases (for example, those regarded as insulting), but to the fact that the entirety of the statement made may be regarded as defama-

[46] CJEU, Case C-324/09, L'Oréal SA and Others v eBay International AG and Others (2011) EU:C:2011:474, para 139.
[47] CJEU, Case C-70/10, Scarlet Extended SA (2011) EU:C:2011:771, para 39.
[48] CJEU, Case C-484/14, Mc Fadden v Sony (2011) EU:C:2016:689, para 101.
[49] CJEU, Case C-18/18, Glawischnig-Piesczek v Facebook (2019) EU:C:2019:821, para 19.

tory.[50] Therefore ensuring the effective protection of a victim's rights requires that an injunction issued by the court covers not only the wording used in the content found to be unlawful, but also "information, the content of which, whilst essentially conveying the same message, is worded slightly differently, because of the words used or their combination".[51] The Court thus interpreted the concept of "information with an equivalent meaning" and considered it permissible for an injunction issued by a national court to also cover the obligation to block that type of information. It pointed out that, in any event, differences in the wording of the content cannot require an online service provider to make an independent assessment of that content, since such an obligation would directly contravene the prohibition laid down in Article 15(1) of the e-Commerce Directive.[52]

The judgment in *Glawischnig-Piesczek v Facebook* thus demonstrates an evolution in the position of the Court, consisting of – at least partially – a departure from the earlier interpretation according to which a general obligation to control information, involving a search for specific content (even if precisely defined – as in *Mc Fadden v Sony*), cannot be reconciled with the prohibition defined in the e-Commerce Directive.

In *Glawischnig-Piesczek v Facebook,* the Court held that a search for content of an equivalent nature does not oblige the hosting provider to make an independent assessment of the information just because it has "recourse to automated search tools and technologies". Such an argument, however, is unconvincing. Both the analysis of copyrighted works and the search for defamatory or libellous content are carried out automatically with the use of similar algorithmic systems. Technically, there is no clear difference between the proactive blocking of distribution of a specific multimedia file (which, in the CJEU's view, is incompatible with EU law) and the blocking of specific content or even information similar to it that defames an individual (which, in the Court's view, can be reconciled with the wording of the e-Commerce Directive). The interpretation presented in the earlier cases is thus different from the position expressed in *Glawischnig-Piesczek v Facebook*, which in turn leads to the conclusion that the Court's line of jurisprudence is evolving in the direction of providing broader protection to individuals against repeated infringements they may experience in cyberspace. In this regard, it is irrelevant whether the service provider was previously notified by the victim or a third party of the publication of the defamatory information.[53]

The Court rightly pointed out that it is unacceptable to limit the liability of service providers only to cases where they fail to respond to reports by the aggrieved party. The aggrieved party does not need to know about the publication of the injurious statement; moreover, it does not need to use the service provider in question or Internet services in general. Regardless of their awareness of the infringement, the mere fact of publication may lead to harm. Significantly, the Court stressed that information that incites hatred or violence is clearly unlawful – which must also be obvious to the service provider, who then should immediately take action to block such content.[54] Significantly, in the case examined here, the service provider (Delfi AS) was a professional entity dealing with the commercial operation of an information portal. By its actions, Delfi AS encouraged the publication of comments which, by increasing the number of page views, ultimately led to an increase in the company's revenue. Thus, in essence the service provider earned money by posting comments and facilitating discussions between users. Although the service provider used automatic content filtering mechanisms, their ineffective operation resulted in the removal of offending comments from the website only after a long period of time.[55]

The *Delfi* judgment has significantly expanded the limits of online service providers' liability – indicating that it is not always possible to exclude their liability solely on the basis that they were not informed about the publication of unlawful content. This is particularly the case when, given the totality of the circumstances of the case and the attitude of the service provider, its conduct can be considered to have gone beyond a "passive, purely technical" role.[56]

The use of insufficient and ineffective content filtering mechanisms by a service provider cannot be regarded as the only premise for holding it liable for the publication of unlawful content. Such a conclusion follows from the judgment in *Pihl v Sweden*, in which the Court found no violation of the Convention.[57] The background of the case was the publication of a comment on a website run by an NGO that accused the applicant of Nazi sympathies. The post was removed immediately when the service provider was informed. In deciding the case, the Court took into account the local scope of the website, the non-commercial manner in which it was operated, and the content of the published comment (which did not contain insults or incitement to violence).[58] The Court's position was further developed in the *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary* case, which explained that excessively stringent content filtering rules, applied in particular to non-commercial websites and not related to the actual damage caused by published content, "may have, directly or indirectly, a chilling effect on the freedom of expression on the Internet."[59]

Moreover, in *Jezior v Poland* the Court emphasised that to require a service provider to assume that "certain unfiltered comments may be unlawful would be tantamount to requiring (…) an excessive and unrealistic degree of foreseeability, which could jeopardise the right to communicate on the internet."[60]

Thus, while it follows from the current case law of the European courts that the "notice and takedown" model cannot be the only way to counteract hate speech, at the same time it appears unfair to shift all responsibility for fighting this phenomenon onto service providers by expecting them to imple-

[50] Glawischnig-Piesczek v Facebook case (n 49), para. 40.
[51] Glawischnig-Piesczek v Facebook case (n 49), para. 41.
[52] Glawischnig-Piesczek v Facebook case (n 49), para. 45.
[53] Delfi AS v Estonia case (n 1), para. 159.
[54] Delfi AS v Estonia case (n 1), para. 115.
[55] Delfi AS v Estonia case (n 1), para. 156.
[56] Delfi AS v Estonia case (n 1), para. 146.
[57] ECtHR, Appl. 74742/14, Pihl v Sweden (2017).
[58] Pihl v Sweden case (n 57), paras. 31-32.
[59] ECtHR, Appl. 22947/13, Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary v. Estonia (2016), para. 86.
[60] ECtHR, Appl. 31955/11, Jezior v Poland (2020), para. 58.

ment technical safeguards that would lead to the complete elimination of violent and defamatory content.

5.2. *Copyright protection*

The provisions of the e-Commerce Directive were adopted in 2000, at a time when the digital services available today did not exist. With the rapid development of global e-services, copyright holders have increasingly pointed out the inadequacy of the measures taken by leading service providers to minimise copyright infringement. In particular, they have argued that algorithmic mechanisms used for identifying copyright infringements are ineffective. They have also pointed out that, in the case of a notification of copyright infringement, service providers have blocked only the content directly challenged in the notification and not all files available as part of a given service with identical content (see the previous comments on the *notice and takedown* mechanism defined in the ECD). Service providers, in turn, have argued that it is not their role to censor the content published by users. In their opinion, in the case of violation of the terms of service and the publication of content infringing copyright, the owner of intellectual property rights should seek compensation directly from the infringer – i.e. from a specific user.

This issue was examined by the Court of Justice in the *Frank Peterson v Google LLC* case.[61] In its ruling, the Court interpreted the notion of *communication to the public* – introduced by the Copyright Directive[62] and of fundamental importance for establishing the service provider's liability for copyright infringement. The Copyright Directive defines authorisation of communicating content to the public as an exclusive right of the author. While making a protected work available via an electronic service (e.g. a file-sharing platform) certainly leads to an infringement of this right, it is not clear who is liable for such act – the service provider or the user who published the specific content.

According to the Court, as a general rule the mere fact that a service provider is aware that there may be unlawfully hosted content on its platform is not sufficient for it to be regarded as allowing internet users access to that content.[63] In this regard the Court stated that a service provider contributes to making publicly making the content if any of the following conditions are met[64]:

First, if it has reliable knowledge of the unlawful character of the content and, despite that fact, does not remove it or block access to it (*cf* the previous analysis of Article 14(1) of the ECD). Secondly, if it, being aware of the possibility of distribution of unlawful content by users, does not implement appropriate technical solutions. Thirdly, when, taking into account the functionality of a given service, it should be deemed that the operator participates in the choice of or supports publishing the illegal content – i.e. when the business model adopted by it encourages users to upload content infringing copyright.

In the Court's view, mere statistical information on the number of files infringing copyright uploaded to a particular service is insufficient to establish that the operator accepts and condones the distribution of the illegal content. Also relevant to that assessment is whether the service provider "put in place the appropriate technological measures that can be expected from a reasonably diligent operator in its situation in order to counter credibly and effectively copyright infringements on that platform".[65] Of course, the requirement to use technological measures does not predetermine the need to implement an algorithmic mechanism leading to preventive filtering of content.

In principle, the judgment in *Frank Peterson v Google LLC* was in line with the expectations of service providers – in particular it did not lead to broadening the scope of their responsibility for content published by users. Although the Court emphasised the need to use technical means – adequate to the capabilities and capacity of a given operator – it did not conclude that these mechanisms must be applied in a preventive manner or that faults in their operation (leading to the publication of content that infringes copyright) make the service provider liable.

In 2019, even before the ruling in *Frank Peterson v Google LLC*, the EU legislature adopted Directive 2019/790 (Copyright in the Digital Services Market Directive, CDSM Directive),[66] one of the objectives of which was to clarify the rules concerning the liability of service providers for making content available online. In this regard the legislature decided to introduce a *lex specialis* model, creating a regime making the general requirements under the ECD more specific. Thus, the Directive does not cover all information society service providers, but only those entities whose "main or one of the main purposes is to store and give the public access to a large amount of copyright-protected works (…) uploaded by its users".[67] Thus, while the liability regime of, for example, hosting providers will continue to be governed by the provisions defined in the e-Commerce Directive, entities that meet the definition of online content providers (e.g. YouTube) are subject to the new provisions arising from Article 17 of the CDSM Directive.

The main change introduced by the provisions of the CDSM Directive is the obligation of service providers to obtain authorisation from copyright holders in order to make copyrighted content publicly available on the Internet. Only if obtaining such permission proves impossible will the service provider be able to exempt itself from liability for copyright infringement, by demonstrating that it made its best effort to prevent access to content that it knew was illegal. In such a case the service provider will not only have to react immediately to the information provided by the authorised entities, but also make every effort to prevent future posting of the disputed content by users.

[61] CJEU, Joined Cases C-682/18 and C-683/18, Frank Peterson v Google LLC and Others (2021) EU:C:2021:503.

[62] 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, OJ of 2001 L 167/10.

[63] Frank Peterson case (n 61), para. 85.

[64] Frank Peterson case (n 61), para. 84.

[65] Frank Peterson case (n 61), para. 102.

[66] 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 of 2019 L 130/92.

[67] Article 2(6) of CDSM Directive.

As a result, the CDSM Directive imposes on providers of online content services an obligation to identify and block access to content which is identical to works claimed by copyright holders. In this regard the Directive uses the concept of due care ("*best efforts*"), that is to say care which characterises the actions taken by a diligent entity to achieve identical results, having regard for the best practices in the relevant sector and observing the principle of proportionality. Only by demonstrating the exercise of due care can the entity be released from liability for publishing content that was previously reported as protected by copyright. At the same time, the legislature did not specify what criteria should be used to determine that the link between the examined work and the protected work is so significant that it should result in blocking its publication.

This led to concerns that the adoption of the new legislation could lead to the so-called "over-blocking" or "collateral censorship" effect.[68] This can occur in the case of a flawed legal model that makes an intermediary providing information, whose identity is easily identifiable, liable for the statements of third parties. In such a case, the intermediary (service provider) may have an interest in eliminating the risk by limiting the free speech of those for whom it could be held liable. This may lead, among other things, to a configuring of the technical means applied that minimises the number of *false negatives* while ignoring the growing rate of *false positives* among the content analysed.

Although in the CDSM Directive the EU legislature repeated the prohibition on establishing a general control mechanism, a prohibition already introduced in the e-Commerce Directive, NGOs and the media pointed out that the new regulations may negatively impact freedom of speech and the right to information.[69] It has been highlighted, *inter alia*, that the CDSM Directive in fact serves to implement into the EU legal order measures similar to the widely criticised ACTA draft agreement.[70] Comparisons between the CDSM Directive and ACTA were further emphasised by the use of terms such as 'censorship machine' in articles and comments on the new Directive.[71] It was also pointed out that there existed a future risk that the private sector, or governments, would decide to expand the use of content filtering mechanisms originally established to prevent copyright infringement.[72] In addition, the Polish Prime Minister, when commenting on the draft Directive, indicated that it "fuels censorship and threatens freedom of expression".[73]

These doubts led Poland to bring an action before the Court of Justice seeking a declaration that Article 17 of the CDSM Directive was invalid insofar as it imposed an obligation on entitites to use their best efforts to prevent access to content in respect of which a well-founded objection had been raised that it infringed copyright.[74]

In his opinion, however, Advocate General Henrik Saugmandsgaard Øe did not share these concerns. First of all, he pointed out that it was clear from the wording of the Directive that its aim was not to establish a system of general censorship. In a detailed analysis of the content of the contested provisions, he aptly pointed out that the system established by the CDSM Directive cannot be used to introduce preventive blocking of all reproductions of protected works – but only of content that is identical to protected works. In this respect, he explained that the identicalness should be obvious in the light of "relevant and necessary information provided by the right holders".[75] Thus he took the view that the role of a service provider is not to decide whether a particular work is sufficiently similar to a protected work, but to rely in this regard on the indications of the right holders. He consequently held that short fragments of works and transformed works cannot be subject to preventive blocking.[76] According to AG Saugmandsgaard Øe, the role of service providers should – in the light of the CDSM Directive – focus on finding identical works and works containing insignificant modifications, irrelevant or indistinguishable to the recipients.[77] In his opinion the detection of such content does not require an independent assessment (defined as a substantive assessment of the content) and can be carried out automatically.

Although AG Saugmandsgaard Øe's position is consistent, it does not clarify all the doubts set out in the Polish application. Undoubtedly, the control model envisaged by the CDSM Directive is intended – at least in part – to shift responsibility for identifying copyright-infringing material from rightholders to providers of online services (content delivery services). It should be noted however that intellectual protection is one of the most dynamically-developing areas of law, in which the interpretation of particular concepts is subject to constant change. The right to quote or the right to re-use may be defined and interpreted differently, depending on the context and circumstances of each case. Authors of online content often experiment with different forms of expression and, from the perspective of a service provider, the line between what is an unlawful use of a protected work and its creative transformation can be much more difficult to draw than Saugmandsgaard Øe's analysis suggests.

[68] See the joint dissenting opinion of Judges Sajó and Tsotsoria presented in the case Delfi AS v Estonia (n 1). The problem of collateral censorship is widely discussed in the United States – see e.g. Felix T Wu, 'Collateral Censorship and the Limits of Intermediary Immunity' (2013) 87 Notre Dame Law Review 293.

[69] *See e.g.* Cory Doctorow, 'The Final Version of the EU's Copyright Directive Is the Worst One Yet', Electronic Frontier Foundation (13 February 2019), <https://cli.re/1Xjo7R> accessed 24 September 2021; also n 7.

[70] Note that dissatisfaction with the draft ACTA agreement led to civil unrest and riots in several member states. See Dave Lee, 'Acta protests: Thousands take to streets across Europe', BBC News (8 March 2021), <https://cli.re/Vbz8mV> accessed 24 September 2021.

[71] 'Censorship machine takes over EU's internet', European Digital Rights Press Release (26 March 2019), <https://cli.re/KaRvKM> accessed 24 September 2021.

[72] Felipe Romero-Moreno, '"Notice and Staydown" and Social Media: Amending Article 13 of the Proposed Directive on Copyright' (2019) 33 International Review of Law, Computers & Technology 187, 204.

[73] Matt Reynolds, 'What is Article 13? The EU's divisive new copyright plan explained', Wired (24 May 2019), <https://cli.re/aD8RKY> accessed 24 September 2021.

[74] See Art. 17(4)(b-c) of the CDSM Directive.

[75] See n 8, para. 205.

[76] See n 8, para. 206.

[77] See n 8, para. 202.

Opponents of the CDSM Directive have postulated that its adoption could have a 'chilling effect' on online service providers.[78] This would consist of an automatic assessment of the published material in such a way as to exclude from publication any content that would entail the risk of a legal dispute (see previous comments on the 'over-blocking effect').[79] Taking into account the huge amount of content published each day on the sites covered by the Directive, and bearing in mind the incremental effect that characterises the functioning of machine learning systems, even a small number of incorrectly rejected materials may over time lead to a censorship system which extends beyond the original purpose of its establishment.

According to AG Saugmandsgaard Øe there is no such risk, because under the CDSM Directive, service providers are not entitled to carry out a substantive assessment of the content published. At the same time however, in defining the obligations of service providers the Advocate General used very vague terms, which can be subject to various interpretations. In practice, it is impossible to assess whether a work is a reproduction of another work without an analysis of it, which goes beyond a comparison of the attributes of the file itself (the length of the recording, size of the file, etc.). Systems which perform such an analysis are based on advanced algorithms that process the entire content of the material and – as previously indicated – build parameters used in further assessment. It should be noted though that this evaluation is carried out automatically and when certain parameters exceed defined thresholds, the publication is blocked. Put simply, the role of the service operator is to set the threshold values of these parameters. The processing itself is carried out automatically and on a mass scale. The algorithm does not understand what "*obvious*" unlawfulness is, or that the unlawful nature of given content is "*manifest*" – criteria that, according to AG Saugmandsgaard Øe, should be taken into account by service providers.[80]

Manifest unlawfulness – as defined by the ECtHR in *Delfi AS v Estonia* in the context of hate speech – is not the same as *manifest* unlawfulness in relation to copyright infringement. Hate speech can be identified algorithmically with relative ease, whereas identifying copyright infringement may in many cases require an understanding of the content of the work being examined in order to assess its apparent similarities.[81]

Furthermore, the arguments presented by AG Saugmandsgaard Øe seem to ignore the fact that in the world of machine learning systems, results are expressed in confidence percentages, and in consequence nothing is *obvious*. It is easy to find examples of how classification systems have malfunctioned, for instance identifying a human in a photo as an ape with a confidence of over >90%.[82]

These reservations also led to doubts about whether the measures introduced by the CDSM Directive did not breach the prohibition under EU law on imposing a general obligation of control on service providers. AG Saugmandsgaard Øe, relying on recent case law from the Court – in particular the *Glawischnig-Piesczek v Facebook* case took the view that the filtering of content, although carried out in a generalised manner (i.e. it concerns all content made available by all users), is in fact of an individual nature since its aim is to identify concrete material which, if published, could lead to an infringement of the law.[83] The argument that a different interpretation would *de facto* make it impossible for the EU legislature to take effective measures against certain forms of online crime should be considered accurate.[84]

With regard to this aspect, AG Saugmandsgaard Øe emphasised the impossibility – stemming from the e-Commerce Directive and the earlier CJEU case law – of requiring a service provider to make an "independent assessment" in identifying unlawful content. It is questionable, though, whether the operation of modern algorithmic systems meet this criterion. The mere fact that content is analysed by an algorithmic system – without the participation of personnel – cannot be equated with a lack of independent assessment. In this context, the term *independent* should be understood as an independent assessment going beyond the instructions received from the copyright owner. AG Saugmandsgaard Øe pointed to the identification of files with a changed playing speed or rotated image as examples where such an independent assessment does not occur.[85] There are many such categories of similarity, and it seems that the very identification of such similarities and the establishment of acceptable thresholds for each of them may easily lead to crossing the line of independent assessment.

Nevertheless, the adoption of the CDSM Directive should not be perceived as a genuine risk that a system of applying preventive censorship generally may be adopted. In the first place, the CDSM Directive *de facto* serves to regulate private-law relationships – in particular the rules governing liability between copyright holders and online content providers. Moreover, it applies only to a particular category of digital service providers. In addition to this, it does not grant public entities any power to block content published online. This does not change the fact that further discussion is necessary on the limits of the liability of entities obliged to apply the CDSM Directive, and that it should take into account the specific nature of the means used by those entities to control the content made available. Therefore, it appears that – irrespective of the

[78] For more on the "notice and staydown" model as proposed in the CDSM Directive, see: Romero-Moreno (n 75).

[79] FedericoFerri, 'The Dark Side(s) of the EU Directive on Copyright and Related Rights in the Digital Single Market' [2020] China-EU Law Journal <http://link.springer.com/10.1007/s12689-020-00089-5> accessed 26 September 2021.

[80] See n 8, para. 198.

[81] In practice, the elimination of hate speech is mostly not done algorithmically – according to Facebook, only 38% of deleted content was identified automatically by algorithms, compared to over 99% for terrorist content. Jason Koebler and Joseph Cox, 'The impossible job: inside Facebook's struggle to moderate two billion people', Vice (2018), <https://cli.re/mrDwrA> accessed 24 September 2021.

[82] Maggie Zhang, 'Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software', Forbes (1 July 2015), <https://cli.re/D3m5Nw> accessed 24 September 2021.

[83] Glawischnig-Piesczek v Facebook case (n 49), paras. 34-35; see also n 8, para. 114.

[84] See n 8, para. 113.

[85] See n 8, para. 202.

final outcome of the *Poland v Parliament and Council* case – the application of content filtering rules on the basis of the CDSM Directive will in the future be the subject of subsequent cases pending before the CJEU.

6. The fight against serious crime

Irrespective of the discussion on the need for and scope of content filtering mechanisms to protect private interests, the EU legal model has also introduced the same measures in the field of criminal law, with a view to facilitating the fight against serious crime. The main difference between the two categories of measures is the role of the public authorities in deciding the scope of information that should be blocked or removed from cyberspace. In the case of the protection of private interests, the powers of public authorities are considerably limited, with service providers taking action at the request of an interested party. In turn, in the case of general security objectives, it is usually the public authorities that initiate an action to remove contested content from online services. The setting up of content filtering mechanisms, even if only partly carried out on the initiative or under the control of the public authorities, requires an analysis of the adequacy of the safeguards put in place to protect against the risks of abuse of power and arbitrary decisions.

The EU legislature has chosen not to adopt a unified legal framework dedicated to preventive content filtering for public purposes. The existing rules are thus scattered, have been adopted at different times, and are based on different organisational and technical measures. In particular, these include:

1) the removal and blocking of websites containing or disseminating child pornography, as established by Directive 2011/93 ;[86]
2) removing or preventing access to terrorist content, as provided for in Directive 2017/541[87] and Regulation 2021/784[88];
3) the use of modern data processing technologies to detect child sexual abuse, arising from Regulation 2021/1232.[89]

At the same time, in a number of specific provisions harmonising the fight against serious crime the EU legislature has not regulated the use of content filtering mechanisms at all. An example is Council Framework Decision 2008/913[90] on combatting certain forms of racism and xenophobia. The issue of insufficient EU measures to combat such phenomena, including hate speech, was identified by the Commission as an area requiring further legislative work.[91]

6.1. *Child sexual abuse (Directive 2011/93)*

Historically, the first regulations that not only allowed – but also defined – an obligation to apply online content filtering measures were the regulations introduced in Article 25 of Directive 2011/93. Member States were obliged to establish national measures ensuring the removal of child pornography content distributed from servers located within their territory. In addition, the Directive also provided for the possibility of introducing mechanisms for blocking sites containing such content located on foreign servers.

The EU legislature has not laid down more detailed procedures for both removing and blocking contested content. In particular, there is no maximum time limit for removing or blocking the material, no proactive obligation for service providers to identify paedophile content, and no sanctions for non-compliance.

As the Commission's report shows, Member States have chosen to meet their obligations in two main ways: by relying either on the "notice and takedown" mechanism under Directive 2000/31 (discussed in the previous sections); or on measures under national criminal law.[92] Regarding the first group, national hotlines have been set up to help monitor paedophile content and report identified cases to service providers. Depending on the specific procedures adopted in each country, information may also be communicated simultaneously to law enforcement authorities. The practice of informing national law enforcement authorities is also applied when prohibited content is found on servers in other countries. In fact, since Member States have a great deal of freedom in transposing the provisions of the Directive, it is difficult to speak of a unified procedure in the event of discovering child abuse content on servers in another Member State. However, the lack of standardisation in this regard has not reduced the effectiveness of the mechanism. The Commission's figures show that 93% of illegal material in the Union and 91% of material hosted abroad was blocked in less than 72 h.[93]

The second mechanism introduced in Directive 2011/93 provides a framework for Member States to take measures to

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

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

[88] 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.

[89] 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 regarding the use of technologies by providers of number-independent interpersonal communications services for the processing of personal and other data for the purpose of combatting online child sexual abuse, OJ 2021 L 274/41.

[90] Council Framework Decision 2008/913/JHA of 28 November 2008 on combatting certain forms and expressions of racism and xenophobia by means of criminal law, OJ 2008 L 328/55.

[91] 'Extension of the list of EU crimes to hate speech and hate crime', European Commission Roadmap (23 February 2021), <https://cli.re/yr8o4n> accessed 24 September 2021.

[92] Report from the Commission to the European Parliament and the Council assessing the implementation of the measures referred to in Article 25 of Directive 2011/93/EU of 13 December 2011 on combatting sexual abuse and sexual exploitation of children and child pornography, European Commission (16 December 2016), COM(2016) 872 final, <https://cli.re/b9221e> accessed 24 September 2021.

[93] See n 92, para. 2.1.2.

block foreign websites that share or disseminate child pornography. Due to its voluntary nature, only about half the Member States have chosen to introduce such national legislation. In the group of countries where such measures have been adopted, blocking is based on either a court order (Greece, Spain and Hungary); on an administrative decision that is a mandatory (France, Italy, Cyprus and Portugal); or one that is voluntary (Ireland, Sweden and the United Kingdom).[94] It is common practice to establish lists of prohibited sites (*black lists*), which are managed by public bodies and made available to internet service providers.[95] Significantly, Article 25 of Directive 2011/93 does not specify the technical means of blocking content – there are no formal obstacles to the establishment of a blocking mechanism implemented not by ISPs but, for example, by providers of mobile devices. It should be noted that the directive *de facto* focuses on the harmonisation of Member States' national laws; it does not introduce separate transnational cooperation mechanisms.

6.2. *Blocking terrorist content (Regulation 2021/784)*

The EU legislature has provided far more detailed requirements for blocking terrorist content. These originally stemmed from Directive 2017/541, which established a regulatory model similar to that of the provisions contained in Directive 2011/93, discussed above. It consists of two related measures: obliging Member States to adopt legislation that ensures the removal of content publicly inciting persons to commit a terrorist offence; and providing for the (optional) establishment of mechanisms to block domestic users' access to such material located on foreign servers.

However, the regulatory model resulting from Directive 2017/541 has been significantly expanded by the adoption of the new Regulation 2021/784. The EU legislature has thus introduced a new mechanism of transnational cooperation – based on the ability of authorised bodies in Member States to issue takedown orders, with the effect of preventing access to contested content in all Member States. It will be possible to address such injunctions only to hosting service providers – which is supposed to increase the speed of reaction and ensure that the offending materials are blocked (or removed) at source. Notably, the entities obliged to apply the Regulation are not providers of electronic communications services, such as e-mail or instant messaging providers.[96] Neither are they IT infrastructure providers (including cloud service providers).

The scope of a takedown order may cover not only content that incites the commission of a terrorist act (which contents were also subject of measures defined in Directive 2017/541), but also materials related to recruitment, training or the provision of instruction, or those that encourage participation in a terrorist group or pose a risk of conducting a terrorist act. Orders issued may therefore cover a wider range of content than national measures adopted in the transposition of the provisions of Directive 2017/541. However, the broad definition of terrorist content raises concerns about the possibility of using this mechanism to remove content only loosely related to the area of public security. As Delbos-Corfield pointed out, "Hungary theorised that you were an enemy of the nation, which is not far from terrorist, when you criticised an element of the government".[97]

One of the more controversial provisions of the Regulation is the obligation to remove the content immediately, and no later than within one hour. Such a short period of time significantly limits the service provider's ability to conduct any analysis, including whether there are any grounds to not execute the order received. Although the EU legislature has indicated that the execution of a removal order must not lead to the blocking of educational or journalistic content, including content that "expresses polemic or controversial views as part of a public debate", this declaration has not been accompanied by any mechanisms disallowing the removal of such publications.[98]

It will be possible to address removal orders both to service providers who have their head office in the same country as the issuing authority as well as to those in any other Member State. For cross-border removal orders, information about its issuance should also be provided to the competent authority of the country where the recipient is established. As a general rule, this authority will have 72 h to verify "whether it [the removal order] seriously or manifestly infringes the Regulation or the fundamental rights enshrined in the Charter". Such verification may also be carried out at the request of a foreign service provider. The provisions of the Regulation further grant both the hosting provider and the content provider the right to an effective remedy both before the courts of the issuing state and before the courts of the Member State where the order has been enforced.

The use of the *1-hour takedown* mechanism raises serious concerns about its compliance with the principles of proportionality and necessity. It is worth recalling that in 2020 the French Constitutional Council challenged the constitutionality of similar national legislation, which also provided for the issuing of administrative orders to remove content related to terrorism or child sexual abuse.[99] In its ruling, the Conseil Constitutionnel stressed that "the one hour period given to the publisher or the obliged entity to withdraw or make inaccessible the targeted content does not allow them to get a ruling from a judge before being forced to withdraw it".[100] The absence of any compulsory judicial review, both at the stage of issuing the order and at the stage of its execution, means that to a large extent the arguments of the French constitutional court can be repeated with respect to the measures established by Regulation 2021/784. Since the Constitutional Council has found similar national provisions to be in breach

[94] See n 92, para. 2.2.

[95] See also previous comments on the limited effectiveness of Poland's gambling blocking mechanism.

[96] *See* recital 14 of the Regulation 2021/784.

[97] Mathieu Pollet, 'EU adopts law giving tech giants one hour to remove terrorist content', EURACTIV (28 April 2021), <https://cli.re/Rwyn73> accessed 24 September 2021.

[98] The Regulation sets out the legal remedies which may be used to republish the contested material. The Regulation does not specify that the use of legal remedies suspends the enforceability of the decision to remove/block the content.

[99] Conseil Constitutionnel, Case 2020-801 (June 18, 2020), FR:CC:2020:2020.801.DC, <https://cli.re/pZr2Jw> accessed on September 24, 2021.

[100] See n 99, para 7.

of the French Basic Law, it seems likely that the Court of Justice will come to the same conclusion when examining the compatibility of the Regulation's provisions with the guarantees defined in the Charter of Fundamental Rights.

Regulation 2021/784 also introduces additional obligations for hosting providers to apply so-called "specific measures". These concern the identification and preventive removal of terrorist content. The application of specific measures is limited to entities that have received a relevant decision from a competent authority. As a rule, the decision as to which specific measures are to be applied remains with the provider of the hosting service, and the Regulation merely sets out examples of the categories of action that may be taken. These include not only the implementation of technical and organisational systems to facilitate content identification and blocking, but also mechanisms for reporting prohibited content and measures to raise user awareness. The European Parliament emphasised that the requirement to take specific measures cannot include "an obligation to use automated tools by the hosting service provider".[101] The aim of this provision is to prevent a competent authority, which has taken a critical view of the measures applied so far by a service provider, from obliging the provider to implement automated content filtering mechanisms.

The legislative work on Regulation 2021/784 was accompanied by intense public debate, during which NGOs in particular pointed out the risks associated with the adoption of the new law.[102] Since the provisions of the Regulation will not be applied until 7 June 2022, the validity of some of the doubts raised cannot be resolved at the present time.

6.3. *Identification of infringements through surveillance measures (Regulation 2021/1232)*

The most recent example of an EU legal provision related to the use of content filtering mechanisms to fight serious crime are the regulations introduced by Regulation 2021/1232. Like Directive 2011/93, this Regulation deals with preventing access to and dissemination of paedophile content. However, while the provisions of the Directive were addressed to Member States and aimed at establishing national measures obliging service providers to remove and block contested content, the purpose of the Regulation is to allow providers of certain electronic communications services to use modern data processing measures to identify cases of child sexual abuse. To this end, the provisions of the Regulation establish an exception to the application of the EU laws on data protection (Regulation 2016/679) and security in electronic communications (Directive 2002/58) by extending the permissible scope of use of data by providers of the so-called number-independent interpersonal communications services (such as operators of instant messaging services like Skype or Viber).[103] Crucially, the purpose of the Regulation is to introduce a legal basis for the use of automated mechanisms for analysing traffic and content data exchanged between users of communication systems and for reporting them to authorised authorities in the event of suspected paedophilic content.[104]

Moreover, although the provisions of the Regulation describe in great detail the obligations of the service provider, indicate a data retention period, strictly limit the scope of permissible processing, and contain a number of other legal safeguards, they do not, in principle, exclude any technical means used to identify illegal material. An analysis of the Regulation's provisions even shows that the legislature has created a framework for service providers to develop and use new ways of user surveillance which would be more effective in detecting infringements.[105]

While there is no doubt that combatting the sexual exploitation of children is imperative, it cannot be ignored that the measures introduced to achieve this goal can easily be applied in other areas in the future. Therefore, when discussing the proportionality of a measure such as that introduced in Regulation 2021/1232, it is worth bearing in mind that even the most legitimate aim cannot justify the introduction of a widespread system of surveillance, not to mention one operated under the supervision of public authorities.

Providing for an exception to EU data protection law does not limit the application of the Charter of Fundamental Rights. Therefore, if the reasoning presented in the Regulation – leading to the possibility of a permanent and generalised monitoring and analysis of all users' communications – were to be accepted, the same argument can be used to demonstrate the necessity and proportionality of a similar measure used for the purposes of fighting terrorism or any other serious crime. In its judgments in *Digital Rights Ireland* and *Tele2 Sverige,* the Court of Justice pointed out the disproportionality of establishing generalised forms of electronic communications monitoring that concerned persons in respect of whom there was no indication, even indirect, of a link with criminal activity.[106] To put it another way: in a democratic state, all citizens cannot be suspected of being criminals, in this case paedophiles.

Furthermore, positing that the adoption of such a measure as the one laid down in the Regulation is necessary for the purposes of combatting the sexual exploitation of children reveals a further inconsistency on the part of the EU legislature. Indeed, it is incomprehensible why it has chosen to create *a voluntary* model in which service providers may (but are not obliged to) monitor content shared between their own users. If the EU legislature – and the Member States – believe that such a measure is necessary, it should be applied compulsorily, under the supervision and close scrutiny of the courts.

[101] Article 5(8) of the Regulation 2021/784.

[102] 'European Parliament confirms new online censorship powers', European Digital Rights (29 April 2021), <https://cli.re/ozbVZD> accessed 24 September 2021; 'Draft EU Regulation on 'Terrorist Content' Online Threatens Rights', Human Rights Watch (16 November 2020), <https://cli.re/KazXme> accessed 24 September 2021.

[103] See recital 7 of the Regulation 2021/1232.

[104] Under Article 3(1)(g)(iii) of Regulation 2021/1232, a report of this type must be confirmed by a human being. A service provider cannot therefore automatically send content classified as child sexual abuse to law enforcement.

[105] Article 3(1)(d) of Regulation 2021/1232.

[106] CJEU, Joined Cases C-203/15 and C-698/15, Tele2 Sverige (2016) EU:C:2016:970, para 105-107.

The scope of interference with the right to privacy envisaged by Regulation 2021/1232, resulting from the possibility of monitoring and processing all correspondence exchanged by all users of modern electronic communications, is unprecedented and unknown in any previous EU legislation. As a result, the issue of whether its establishment leads to a violation of the essence of the right to privacy must be assessed. It follows from the well-established case law of the CJEU that a measure which introduces such extensive and profound restrictions that it deprives the essence of a right of its meaning cannot be considered proportionate.[107] Put differently, an *exception* and *derogation* cannot become a *norm*.[108] Significantly, the Court of Justice has also applied the test of respect for the "essence of fundamental rights" in its case law to the protection of personal data and the use of generalised forms of electronic surveillance.[109]

Regulation 2021/1232 could constitute a basis for the introduction of many modern forms of user surveillance – including those using machine learning systems. Electronic communications services that could use new surveillance mechanisms are also provided by major technology companies such as Google and Apple. As a result, Apple's recent proposal to build into mobile devices a mechanism to monitor user activity for illegal content is, surprisingly, in line with the evolution of EU law. If providers of electronic communication services can monitor their users in order to identify unlawful material (in this case related to child sexual abuse), the effectiveness of this measure will only increase if they can use all the data to which they have access – thus in the case of Apple not only the information sent in the iMessage application but also all other information available in the entire ecosystem of its products.[110] This, in turn, opens up the possibility of subjecting users to total surveillance by digital service providers, which in turn obviously raises doubts about the compliance of the framework introduced in Regulation 2021/1232 with guarantees stemming from the Charter of Fundamental Rights.

7. The Digital Service Act as an update to the regulatory framework

Recognising the need to reform the e-Commerce Directive, the Commission presented a draft regulation on digital services (Digital Services Act, DSA)[111] in 2020. The new regulation is aimed at providing an updated regulatory framework for the operation of digital services, while maintaining "the core principles of the liability regime and the prohibition of general monitoring".[112] Therefore, although Articles 14–15 of the e-Commerce Directive, discussed in this article, will be repealed upon the entry into force of the DSA, they will be replaced by similar provisions contained in the new regulation. Thus, Art. 14(1) of the ECD – which forms the basis for the notice and takedown model – will be replaced by an identical provision in Article 5(1) of the DSA. In turn, Article 7 of the DSA reproduces the prohibition expressed in Article 15 of the ECD on obliging a service provider to "actively to seek facts or circumstances indicating illegal activity". The amendments proposed in the DSA take into account the interpretations contained in the recent case law of the CJEU, which will, at least in part, facilitate the application of the EU provisions regulating the operation of digital services.[113] Also noteworthy is the introduction into the DSA of an explicit obligation for service providers to apply the service's terms and conditions in a way that respects the "fundamental rights of the recipients of the service as enshrined in the Charter".[114] Although the final shape of the DSA is not yet known, it is expected that the adoption of this regulation will also result in the introduction of new mechanisms strengthening the rights of individuals and ensuring transparency of actions taken by service providers.

However, the entry into force of the DSA will not allay the doubts presented in this article concerning proactive (automatic) content filtering obligations. Firstly, although the regulation introduces numerous changes to the notice and takedown process (in the DSA called "notice and action"), it only slightly affects the current legal framework for preventive content filtering.[115] Secondly, the regulation will not impact the application of other EU acts that have been discussed – such as the CDSM Directive, Regulation 2021/784 or Regulation 2021/1232.

The proposal to transfer the issue of the liability of a digital service provider to a separate regulation, i.e. the DSA, is certainly appropriate. It would seem prudent however that the draft Regulation be supplemented with more detailed obligations related to the use of automatic content filtering measures. In this way the detailed provisions that are currently regulated separately in several directives, regulations and framework decisions could be combined – at least partially – in a single act. Moreover, it would also be possible to establish a common minimum standard of legal safeguards that takes into account the risks associated with the usage and application of such measures. The current draft DSA is in reality no more than a modernisation of part of the e-Commerce Directive. This may not be enough to achieve the ambitious objectives the Commission has set for the new regulation. It should not be expected that the finalised DSA will also be

[107] Koen Lenaerts, 'Limits on Limitations: The Essence of Fundamental Rights in the EU' (2019) 20 German Law Journal 779.
[108] Tele2 Sverige case, para. 104.
[109] Maja Brkan, 'The Essence of the Fundamental Rights to Privacy and Data Protection: Finding the Way Through the Maze of the CJEU's Constitutional Reasoning' (2019) 20 German Law Journal 864.
[110] Stephen Nellis and Joseph Menn, 'Apple says photos in iCloud will be checked by child abuse detection system', Reuters (9 August 2021), <https://cli.re/mrDEkx> accessed 24 September 2021.
[111] 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, European Commission (15.12.2020), COM(2020) 825 final.
[112] EC's draft of the DSA, n. 111, sec. I.
[113] Including the *L'Oréal SA, Scarlet Extended SA, McFadden v. Sony* and *Glawischnig-Piesczek* cases, discussed in previous sections.
[114] Art. 12(2) of the draft DSA, n. 111.
[115] Herbert Zech, 'General and specific monitoring obligations in the Digital Services Act: Observations regarding machine filters from a private lawyer's perspective', VerfBlog (2 September 2021), <https://cli.re/d3zBkz> accessed 24 September 2021.

'future-proof'[116] if it duplicates regulations such as the current Article 15 of the ECD, the application of which already today seems difficult, if not impossible, to reconcile with the detailed provisions laid down in the CDSM Directive or Regulations 2021/784 or 2021/1232.

Moreover, even following the entry into force of the DSA, individual regulations will still partly overlap and, as they may concern the same obliged entities, will make it more difficult to build a coherent model of compliance with all the applicable legal requirements. Such a complicated regulatory model lacks transparency not only for obliged entities, but also for users of online services. Unfortunately, it is unlikely that the entry into force of the DSA will significantly change this assessment. Suffice it to point out that since the publication of the first draft of the DSA, the EU legislature has already adopted two new regulations establishing automatic mechanisms for online content filtering.

8. Summary

The need to ensure adequate protection of the rights of individuals, as well as to achieve general security objectives, is leading increasingly to the establishment of rules that allow the use of content filtering mechanisms. The e-Commerce Directive, which has been in force for more than 20 years, introduced into the EU legal model the possibility for digital service providers to avoid liability if they are not aware of illegal content stored or uploaded by users. The application of these provisions was unchallenged for years, but with the transformation of the entire digital services market and the emergence of disturbing trends – such as the exponential increase in hate speech – they needed to be clarified. As a result, the Court of Justice has provided an interpretation that shows the need for service providers to move beyond the passive 'notice and takedown' model and adopt a more proactive stance, including the use of automatic systems to identify unlawful content.

However, the increasing emphasis on preventive content filtering as a condition for digital service providers to be exempted from liability is leading to the gradual erosion of another legal mechanism established by EU law – namely the prohibition on imposing a general obligation of control. Although the Court of Justice seems to recognise this problem, the interpretation of EU law it has put forward in its recent case law does not sufficiently take into account the specificities of algorithmic systems used to filter content.

Notwithstanding the evolution of the case law of the European courts, the EU legislature is also taking its own action. In recent years, several legislative measures have been introduced which directly address the issue of preventive content filtering. With the entry into force of Regulation 2021/784, a 1-hour takedown – a previously unknown mechanism – appeared in the EU legal order. Moreover, the latest legislation also creates a framework for the use of modern content filtering systems to detect previously unknown infringements and report them to law enforcement authorities.

It may also be reasonably assumed that in the coming years both the EU legislature and Member States will introduce more *lex specialis* regulations, establishing new obligations and bans on the distribution of digital content. An example is the recent initiative of the European Parliament to introduce mechanisms for blocking illegal online sports broadcasts.[117]

As a result, the need to establish a separate EU legal regulation setting out fundamental principles for the application of automatic content filtering is becoming more and more apparent. These regulations, first of all, should set the admissibility criteria for establishing such mechanisms so as to ensure their compliance with the standards of human rights protection. In addition, they should specify the minimum legal safeguards that should be established to minimize the risk that content filtering mechanisms will prevent the free dissemination of views and ideas. The introduction of a regulation providing a legal framework for all online content filtering measures applied in the EU would also make it possible to ensure the coherence of the legal model if it became necessary in the future to establish more far-reaching interference measures than those currently applied.

Moreover, the proposed regulation should not only define the obligations of service providers, but also more precisely define the obligations of public entities in this process. The currently-functioning regulatory model is not only fragmented and inconsistent, but is mostly based on obligations imposed only on digital service providers – transferring to them a significant part of the risks related to the distribution of illegal or harmful online content. Drawing on experiences from other areas regulated by EU law, including also the case of content filtering mechanisms, it is reasonable to propose the creation of independent supervisory authorities (at the national or EU level) and to equip them with control and advisory powers regarding the compliance of the mechanisms applied with legal requirements. Moreover, such bodies could also be responsible for defining detailed standards or guidelines and approving the certification framework used to confirm that the content filtering mechanisms used by individual providers are reliable and transparent.

This proposal differs significantly from the current direction of the evolution of the relevant regulations – it is based on the establishment of a coherent regulatory system, supervised and coordinated by appropriate authorities. Its implementation would at the same time serve to strengthen the effectiveness of the protection of digital service users and would reduce the business risks of service providers. Additionally, it would eliminate another risk that has recently been observed, namely an attempt to control service providers' content filtering operations by individual Member States. It is worth bear-

[116] Dita Charanzová, 'How to make Digital Services Act future-proof?', EURACTIV (19 July 2021), <https://cli.re/9DkP87> accessed 24 September 2021.

[117] European Parliament resolution of 19 May 2021 with recommendations to the Commission on the challenges of sports event organisers in the digital environment, P9_TA(2021)0236.

ing in mind that populist governments – such as in Poland – have recently repeatedly announced[118] the need to introduce national regulations allowing for the imposition of financial penalties on service providers for defective (in their opinion) instances of content filtering. The approval of this scenario would lead to further fragmentation of EU regulations in the field of digital services.

Declaration of Competing Interest

None.

Data Availability

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

[118] 'Profil Konfederacji zawieszony. Ziobro proponuje Lex Facebook' ['Confederation profile suspended. Ziobro proposes Lex Facebook'], Rzeczpospolita (7 January 2022), <https://cli.re/vd71po> accessed 18 March 2022.